The international journal of science / 21 November 2019

index
Collaboration & big science

nature

WAVES OF MUTATION

Barcode system images evolutionary dynamics of laboratory yeast

Under inspection
Why some research institutes are screening their scientists' papers

Cosmic explosions
Extremely high-energy emissions from two γ-ray bursts

Feeling the force
Virtual and augmented reality add touch to their sensory repertoire

Vol. 575, No. 7783
nature.com

The international journal of science / 21 November 2019

nature

Unite against this attack on scientific evidence

US environment agency must desist from a course that could harm the health of people and the planet.

The US Environmental Protection Agency (EPA) last week surpassed its own recent record of getting publicity for the wrong reasons.

The New York Times revealed that the agency's leadership is still actively discussing a rule that would require scientists to supply it with the raw data for studies if the findings are to be taken into consideration in the drafting of environmental regulations (see p. 420). The EPA announced its desire for such a rule, which it is calling Strengthening Transparency in Regulatory Science, in April 2018. It is needed, the EPA says, so that the agency can independently reanalyse and revalidate scientific data and models. The EPA says that it will not recognize studies unless scientists agree to supply such data.

Let us consider the implications of such a rule, were it to be adopted. Many of the data that underpin public-health and environmental studies include information about people who will not have consented to disclosing their confidential data, including where they live; their travel habits; their age and gender identity; and the state of their health.

Many such data were integral to the Six Cities study, published in 1993 by what was then the Harvard School of Public Health in Boston, Massachusetts. This work revealed that people living in polluted cities have shorter lives than people in cleaner cities (D. W. Dockery *et al. N. Engl. J. Med.* **329**, 1753–1759; 1993). The results of the Six Cities study led directly to the imposition of life-saving limits on fine particulate matter from emissions. But this research would have been inadmissible under the EPA's proposed rule.

So the question has to be asked: is there a problem in how science is assessed that needs fixing? Why would the EPA wish to create a rule that could risk worsening human and planetary health? Why would the EPA's leaders choose to override their own science advisers, who questioned the rule? Even the US Department of Defense said in August 2018 that the absence of underlying data "should not impede the use of otherwise high-quality studies".

Answers might be found by considering the rule in the context of the wider actions of the administration of President Donald Trump on the environment so far. Whether it's cancelling the Clean Power Plan — the previous administration's signature climate policy — withdrawing from the Paris climate agreement, weakening fuel-efficiency standards or cutting back on environmental research, the US administration is choosing to act against the consensus of the scientific community. The Strengthening Transparency in Regulatory Science rule needs to be viewed against the backdrop of this reality.

> **"Why would the EPA choose to override its own science advisers?"**

The EPA has denied that the rule would be applied retrospectively, or to existing environmental standards. That might be true up to a point. But what would happen when existing standards needed to be reviewed — as most periodically are? Would the rule be applied because the reviewed version would be a future standard? And, if so, would any science — new or old — become inadmissible unless the underlying data and models were supplied? The EPA has yet to clarify what would happen in such a scenario, but last week's revelations had the result of once again uniting the United States' scientific, medical and health communities, and culminated in a crescendo of opposition.

The scale and volume of this response should rattle the EPA's leadership, and the response needs to get bigger and louder still. That will compel the agency to conduct more of the discussion around its rule in public, as it is now doing. Institutions and individuals must redouble their efforts. They must write to their elected representatives to call out this attempt to undermine accepted scientific practice in public-health and environmental standards.

The EPA was created to protect the nation's environment. As it approaches its 50th birthday next year, it must not be allowed to continue on a course of action that will weaken its ability to fulfil that role.

Germline editing needs one message

Science academies and the World Health Organization must speak with one voice on human germline genome editing.

A year ago this week, geneticist He Jiankui made the shocking announcement of the birth of twin girls in China whose genomes had been edited to prevent HIV infection. Undeterred by the global opprobrium heaped on He, Russia's Denis Rebrikov told *Nature* last month about more experiments involving gene editing of human eggs, to help deaf couples give birth to children who would lack the genetic mutation carried by their parents that impairs hearing.

At the same time, every month seems to bring another gene-editing advance. The latest tool, a precision 'search and replace' technique called prime editing, was described in *Nature* last month by David Liu at the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, and his colleagues (A. V. Anzalone *et al. Nature* http://doi.org/dczp; 2019). Randall Platt at the Swiss Federal Institute of Technology (ETH) in Basel called it a "giant leap" towards the goal of making specific changes to the blueprint of life.

The speed of technological advance, coupled with some

nature

scientists' determination to press ahead with editing human germline cells – eggs, sperm and embryonic cells – has been sounding alarm bells for nearly five years. Editing could produce unpredictable changes that an individual's descendants will inherit – with potentially wide-reaching societal implications. Academies, governments and ethicists have been considering how to regulate this. But the manner in which it is being done is suboptimal.

In 2018, the World Health Organization (WHO) set up an independent expert panel to advise on the oversight and governance of human genome editing. A separate international commission on the clinical use of human germline genome editing gathered for its second meeting in London last week. This commission was established by the US National Academy of Science, the US National Academy of Medicine and Britain's Royal Society, to recommend standards and criteria for germline genome editing. Both will report next year, and the commission's report will feed into the WHO process.

But the WHO panel has already recommended setting up a public registry for genome-editing experiments. It has also made an interim recommendation that "it would be irresponsible at this time for anyone to proceed with clinical applications of human germline genome editing", which has been accepted by the agency's leadership. The international commission has yet to say what it thinks, but it would make little sense for it to disagree.

It isn't entirely clear why separate initiatives are needed, and it is unfortunate that representatives of people with disabilities are not part of the decision-making process. However, it isn't too late to rectify these issues, and the two initiatives must, in the end, converge.

There are very real risks that unregulated clinics claiming to be able to eliminate inherited conditions will use untested, possibly harmful procedures. A sure-fire way to give such clinics the green light is an absence of agreed global standards. When the two groups report next year, they must speak with one voice and have more inclusive representation.

A shock to the system

California's universities must help to design and build a clean and resilient power grid.

Confusion reigned the first time that the University of California, Berkeley, lost its connection to the city's electricity grid, on 9 and 10 October. Campus officials were unable to say how long the university's power plant could provide emergency electricity for crucial facilities – such as freezers containing valuable research specimens. Some scientists didn't even know which electric plugs to use to access back-up power. As a precaution, researchers packed freezers with dry ice, and some sent their most important samples to other institutions.

This chain of events can be traced back to last November, when a faulty transmission line sparked the deadliest wildfire in California's history. The Camp Fire tore through the town of Paradise, killed 86 people and levelled thousands of homes and businesses.

Faced with an estimated US$30 billion in insurance claims from that fire and others in 2017, the state's largest utility provider, San Francisco-based Pacific Gas and Electric Company (PG&E), filed for bankruptcy in January. Then, when hot, dry winds raised the fire danger in early October, the company cited legitimate liability concerns and shut down major sections of the electricity grid to prevent more blazes from breaking out.

Evidence that global warming is promoting more frequent and severe wildfires has been mounting for decades, and the fact that electrical equipment can start fires, and contribute to their spread, is hardly news. But few could have predicted that vast stretches of California – the world's fifth-largest economy and a global hub for research and innovation – would be paralysed by a combination of wildfire and electricity blackouts.

> **"A safe, clean, efficient, resilient grid has to be a shared responsibility, and not something for politicians alone to fix."**

Safeguarding lives and habitats from these catastrophes has to be the top priority for the state's decision makers. Solutions for upgrading the grid range from the obvious to the technological. Electrical equipment should be kept clear of vegetation, with power lines buried underground, where feasible. Cameras, sensors and other systems could allow grid operators to detect and isolate problems with speed and precision. There are also measures that Berkeley and other institutions can take, such as reducing their energy demands and allocating limited emergency power to only the most urgent needs.

At the same time, California's research and technology institutions, and its decision makers, could harness more of the state's considerable research muscle in energy and energy policy to address the bigger picture: creating a more resilient, cleaner grid for the whole state.

Researchers at Berkeley and elsewhere have spent years developing smart-grid technologies that allow more control of where electricity goes and when. Economists are calculating the costs and benefits of different kinds of energy infrastructure, such as installing solar panels, or using fuel cells powered by renewably produced hydrogen.

More of this pioneering work should be deployed to solve problems in the institutions' home state. Like the back-up power system that Berkeley used when the grid failed, a wider network of increasingly smaller grids that can be isolated or boosted as needed might be the future.

California's fires are now a chronic problem. A safe, clean, efficient and resilient grid has to be a shared responsibility, and not something for politicians alone to fix. The state's dynamic research, technology and innovation communities must step up to solve the problems in their individual organizations and at the same time craft wider solutions that help California – along with regions worldwide – adapt to our thirst for more energy in an increasingly warmer world.

World view

By Jerry Ravetz

Stop the science training that demands 'don't ask'

It's time to trust students to handle doubt and diversity in science, says Jerry Ravetz.

As a child, I realized that my parents spoke in Yiddish when they didn't want me to know what they were talking about, so I became aware that some knowledge was intended only for grown-ups – don't ask. In college, I was taught an elegant theory of chemical combination based on excess electrons going into holes in the orbital shell of a neighbouring atom. But what about diatomic compounds like oxygen gas? Don't ask; students aren't ready to know. In physics, I learnt that Newton's second law of motion is not an empirical, approximate relation such as Boyle's and Hooke's laws, and instead has a universal application; but what about the science of statics, in which forces are balanced and there is no acceleration? Don't ask. Mere students are not worthy of an answer. Yet when I was moonlighting in the social sciences and humanities, I found my questions and opinions were respected, even if only as part of my learning experience.

Observant students will notice that social problems surrounding science are seldom mentioned in official curricula. And now, these pupils are starting to act. They have shamed their seniors into including more diverse contributors as faculty members and role models. Young scholars insolently ask their superiors why they fail to address the extinction crises elucidated by their research. Such subversions are reminiscent of the mass-produced heretical pamphlets circulated by Martin Luther's supporters at the start of the Protestant Reformation in sixteenth-century Europe. The inherited authoritarian political structures of science education are becoming brittle – but still remain largely unchanged from my own school days.

The philosopher Thomas Kuhn once compared taught science to orthodox theology. A narrow, rigid education does not prepare anyone for the complexities of scientific research, applications and policy. If we discourage students from inquiring into the real nature of scientific truths, or exploring how society shapes the questions that researchers ask, how can we prepare them to maintain public trust in science in our 'post-truth' world? Diversity and doubt produce creativity; we must make room for them, and stop funnelling future scientists into narrow specialties that value technique over thought.

In the 1990s, Silvio Funtowicz, a philosopher of science, and I developed the concept of 'post-normal science', building on the Kuhnian terms 'normal' and 'revolutionary' science. It outlines how to use science in a society confronted with high-stakes decisions, where both facts and values are uncertain; it requires drawing on a broad community with broad inquiries. Suppressing questions from budding scientists is sure to suppress promising ideas and solutions.

> "Stop funnelling future scientists into narrow specialties that value technique over thought."

Jerry Ravetz is an associate fellow at the Institute for Science, Innovation and Society, University of Oxford, UK.
e-mail: jerome.ravetz@gmail.com

As a nonagenarian and former historian of science, I know that even foundational building blocks can be questioned. The unifying patterns of the periodic table are now seen, under closer scrutiny, to be riddled with anomalies and paradoxes (E. Scerri *Nature* **565**, 557–559; 2019). Some scientists now wonder whether the concept of biological 'species' contributes more confusion than insight, and whether it should therefore be abandoned (see go.nature.com/2offaav). However, such a decision would affect conservation policy, in which identification of endangered species is crucial – so it is not just an issue for basic science.

Science students generally remain unaware that concepts such as elements and species are contested or are even contestable. In school, college and beyond, curricula highlight the technical and hide the reflective. Public arguments among scientists often presume that every problem has just one solution. When they were students, these researchers had never learnt that they have a right to be wrong.

And when scientists advise on policy, they are pressured to become attached to official stances on issues, or to shun the responsibility entirely. They then find it difficult to resist dismissing all critics as cranks or 'denialists', whose rejection of 'facts' is a sign of their depravity. (To be sure, much of science denial is cynical and self-serving.)

Nonetheless, vacillating advice on complex issues, most obviously nutrition, should be a warning that, from a future perspective, today's total scientific consensus on some policy issue might have been the result of obduracy, a conflict of interest or worse.

Trust in established science will not be protected by exhortations, denunciations and absolutism. Just as a healthy democracy accommodates dissent and dissonance, the collective consciousness of science would do well to embrace doubt and diversity. This could start with teaching science as a great, flawed, ongoing human achievement, rather than as a collection of cut-and-dried eternal truths. There is plenty of material for such a Socratic education in science: physics and cosmology now enjoy creative ignorance; the digital and life sciences abound in moral mazes; and environmental and sustainability sciences demand recognition of complexities. The established 'facts' can function as tools for ongoing dialogues.

I recall a legendary chemistry professor who was inept at getting classroom demonstrations to work – but discussing what went wrong helped his students to thrive. A mathematician friend ran his classes like those in an Athenian agora: pupils discussed every statement in the textbook until all were satisfied. They did very well in exams, and taught themselves when he was absent. Treating people at all levels as committed thinkers, whose asking teaches us all, is the key to tackling the challenges to science in the post-trust age.

ALISON RYDE

News in brief

ENVIRONMENT AGENCY PUSHES TO RESTRICT DATA USE

Scientists are alarmed about the expansion of a proposed rule that would limit which studies the US Environmental Protection Agency (EPA) can use to develop health and environmental regulations.

The supplemental rule builds on a controversial proposal released last year that would prevent the EPA from considering research unless the underlying data are publicly available, according to a leaked draft reported by *The New York Times* on 11 November. Critics of the original proposal feared that it would prevent consideration of research, such as epidemiological studies, based on confidential health data.

That proposal would have applied to a restricted number of studies. But scientists say the leaked supplement is worse because it would expand the rule to cover almost any kind of research. The text also suggests that the rule could apply to data regardless of when they were generated, potentially affecting the agency's consideration of previously published studies.

If the final rule looks like the leaked proposal, "it will fundamentally change the way EPA uses science to make public-health decisions — to the detriment of public health", says Veena Singla, a public-policy and health researcher at the University of California, San Francisco.

EPA officials stressed that the final proposal under review at the White House is different from the leaked draft. The agency must publish the final text and accept public comments before it can finalize the supplemental rule.

VIOLENCE IN HONG KONG DISRUPTS RESEARCH

Three universities in Hong Kong have cancelled classes on campus for the rest of the term after violent clashes between police and protesters erupted in the grounds. Another four universities have also cancelled classes — in some cases for the rest of the year — over safety concerns. And staff at most of the institutions have been told to stay away for several days.

Images show some protesters carrying bows and arrows — one police officer was reportedly hit in the leg with an arrow.

The clashes are the latest flare up in Hong Kong, and follow six months of street protests. These started in June against an extradition bill that would have allowed people to be sent from the territory to mainland China to stand trial or serve criminal sentences.

The protests on campuses are also disrupting research, and some scientists fear that this could dissuade academics from coming to Hong Kong.

Michael Chan, a chemist at the Chinese University of Hong Kong, says he has been unable to access his lab to check on mouse experiments.

Jianhua Zhang, the dean of science at Hong Kong Baptist University, worries that the ongoing protests will have a wider effect on academia. "I anticipate that people will be reluctant to take offers to work with us," he says.

BYE BYE, RYUGU! CRAFT LEAVES ASTEROID

The Hayabusa2 spacecraft is heading home after performing a series of risky and unprecedented manoeuvres on its six-year mission to asteroid Ryugu (**pictured**).

The Japan Aerospace Exploration Agency's (JAXA's) probe gently fired its thrusters at 10:05 a.m. Japan Standard Time on 13 November, moving away from the asteroid at a speed of less than 10 centimetres per second. From 10 December, the probe will start to use its ion engines to propel its journey back to Earth, where it is due to arrive at the end of 2020. A re-entry capsule will deliver its samples to the surface.

Hayabusa2 was launched in late 2014, and arrived at Ryugu in June 2018. It is the first mission to release landers onto the surface of an asteroid; the first to collect a sample from a 'dark' asteroid's surface; and, after bombarding the surface to create a crater, the first to collect a sample of an asteroid's subsurface material.

Just one kilometre wide and shaped like a spinning top, Ryugu is an unusually dark body, probably the result of having a high concentration of carbon. Initial studies based on Hayabusa2's data suggest that Ryugu formed from the debris of an impact between two larger Solar System bodies.

Image story

Bush fires wreak havoc in eastern Australia

L TO R: KIN CHEUNG/AP/SHUTTERSTOCK; JAXA, CHIBA INSTITUTE OF TECHNOLOGY & COLLABORATORS; DARREN PATEMAN/EPA-EFE/SHUTTERSTOCK; DREW ANGERER/GETTY

Firefighters near the town of Nabiac in eastern Australia have been confronting a huge wildfire that has burnt more than 30,000 hectares over the past 10 days. Several hundred conflagrations have burnt more than one million hectares, destroyed more than 450 homes and killed 4 people across the state of New South Wales since 8 November. More than 50 fires were still burning as *Nature* went to press.

Scientists have forecast particularly severe conditions for bush fires this season because large parts of the country are in drought. Climate change is also making fire conditions more frequent and severe.

On 12 November, hot, dry and windy conditions prompted a 'catastrophic' fire warning for vast areas of New South Wales, including Sydney and its surrounding areas; blazes that ignite during these conditions are likely to spread out of control quickly and houses are unlikely to survive. It is the first time that the catastrophic fire rating has been issued for Sydney since new ratings were introduced in 2009.

FEARS OF FOREIGN INTERFERENCE PROMPT UNIVERSITY GUIDELINES

New guidelines will help Australian universities to protect themselves against foreign interference, says the country's government. The advice follows concerns that foreign groups or authorities, such as the government of China, might be seeking to instigate campus activities that are against Australia's interests.

The guidelines, released on 14 November, were developed by the University Foreign Interference Taskforce, which includes representatives from universities, national security agencies and the education department.

Education minister Dan Tehan, who set up the task force in August, said foreign-interference threats against Australia, including its universities, had reached "unprecedented levels", but gave no details at a press briefing.

The guidelines advise universities to undertake due diligence before entering into research or other collaborations with international partners, and to implement robust cybersecurity strategies.

In late 2018 and early 2019, the Australian National University in Canberra experienced significant data breaches, in which hackers accessed 19 years' worth of personal data from the university's network. Media reports have suggested that the hack was perpetrated from China, but the Australian government says the attack has not been attributed to any one country.

Politicians and academics have also raised concerns about some artificial-intelligence projects involving Chinese universities and Australian researchers.

GOOGLE HEALTH-DATA SCANDAL SPOOKS RESEARCHERS

Google and one of the largest health-care networks in the United States are embroiled in a data-privacy controversy that researchers fear could jeopardize public trust in data-sharing practices and, potentially, academic studies.

At issue is an agreement, dubbed Project Nightingale, that gives Google access to the health-care information, including names and other identifiable data, of tens of millions of people without their knowledge. The people were treated at facilities run by the health network Ascension.

Google says that the project, first reported in *The Wall Street Journal* on 11 November, is meant to develop technology that would enable Ascension to deliver improved health care.

Both companies say that they abided by US laws to protect health-care information. But the US Department of Health and Human Services says it is now looking into "this mass collection of individuals' medical records with respect to the implications for patient privacy". Researchers worry that the revelations will undermine trust in studies more broadly. "With these incidents, we undermine public trust to this whole enterprise," warns Effy Vayena, a bioethicist at the Swiss Federal Institute of Technology in Zurich. "At some point, all of the research will get a bad name."

News in focus

PHOTOGRAPHER:REDDIE/GETTY

Waterholes visited by the endangered Gouldian finch contained trace DNA that allowed scientists to detect the bird's presence.

RARE BIRD'S DETECTION HIGHLIGHTS PROMISE OF 'ENVIRONMENTAL DNA'

Researchers are increasingly using traces of genetic material in soil, water or ice to track rare and endangered species.

By Dyani Lewis

DNA gathered from remote waterholes in northern Australia has been used to detect a rare bird in the wild[1] for the first time. The result is the latest milestone in the rapidly maturing science of environmental DNA, in which traces of genetic material from soil, water or ice are used to reveal the presence of plants and animals.

In a study published on 14 November, a team in Australia reports that genetic material collected from waterholes showed that Gouldian finches (*Erythrura gouldiae*) had visited them in the previous 48 hours. Rangers also confirmed the species' presence at the locations.

Scientists have been using environmental DNA (eDNA) analysis for about 15 years, for purposes including tracking rare or elusive aquatic species, such as the great crested newt (*Triturus cristatus*) in the United Kingdom[2]. And in the past few years, the technique has increasingly been used to identify mammals, insects — and now birds — that live on land.

Testing for eDNA is often safer — for both animals and researchers — more cost-effective and, in some cases, more accurate and sensitive than conventional methods used to pinpoint rare and endangered species, scientists say. This is prompting regulatory agencies in a number of countries to adopt the technology to locate creatures, such as the endangered Canada lynx (*Lynx canadensis*) in the United States, or to monitor for invasive species.

But the technique is yet to convince some scientists, who say eDNA results aren't robust enough to be used as the sole basis for making environment-management decisions that can have legal implications for governments and land owners.

Early studies that used eDNA to pinpoint specific species were criticized because of the potential for improper handling of samples to cause cross-contamination, leading to false-positive results. Scientists using the method are detecting only trace amounts of genetic material, so even minute amounts of contamination can taint the results. But

proponents of the field say that the recent adoption of rigorous protocols that avoid or detect contamination have largely addressed such issues.

The first study to show that large-bodied animals and plants drop enough DNA into their environment – through defecation and shedding cells – to be detected[3] was published in 2003. Five years later, another team showed that DNA in pond water could be used to detect the invasive American bullfrog (*Rana catesbeiana*)[4]. Most such studies gather genetic material from aquatic environments because DNA disperses and remains free-floating in water, and can be detected in trace amounts.

Massive time savings

Around 2014, Michael Schwartz, who heads up the US Forest Service's National Genomics Center for Wildlife and Fish Conservation in Missoula, Montana, and his team used eDNA to detect the endangered and hard-to-monitor bull trout (*Salvelinus confluentus*). The researchers initially analysed 124 water samples from waterways across Montana[5], amassing a volume of data equivalent to that collected over the previous 15 years through conventional surveys that used electrofishing, a method that is risky for people and fish, in which a current is run through the water to attract and then net fish. "We were able to do that in eight days," Schwartz says. "We have estimated that it is about two to ten times faster and two to five times more cost-effective to use eDNA compared to electrofishing."

Earlier this year, Schwartz's team published results showing that DNA left in snow tracks or in snow near camera traps could be used to identify the presence of Canada lynx and wolverine (*Gulo gulo*) in Montana, and a small carnivorous mammal called the fisher (*Pekania pennanti*) in Idaho[6]. Conventional methods for detecting the presence of land animals typically involve time-consuming surveys to identify an animal by its tracks alone, or from scat.

In another case, eDNA was more sensitive than conventional methods. When a camera trap image was unable to clearly identify what looked to be a Canada lynx in an area where its presence was unknown to rangers, eDNA extracted from the snow confirmed that the creature was indeed a lynx, says Schwartz.

In some cases, eDNA analyses are being used to enforce policy. In 2014, the UK government approved the use of eDNA analysis for detecting the endangered great crested newt in land-use surveys that are required by law.

With a burgeoning market for eDNA analyses, dozens of companies now offer genetic tests for detecting rare species.

To reduce problems such as false positives that plagued the field in its early days, there are now standard methods for handling samples and detecting contamination, says Florian Leese, an aquatic ecologist at the University of Duisburg–Essen in Germany. Adequate sampling, sterile equipment and experimental controls can all help to guard against contamination. DNAqua-Net, a European-based network of researchers who work with industry bodies and regulatory agencies, is developing best-practice guidelines on how to design and validate tests for individual species and to define the amount of DNA needed to be sure a test returns a genuine positive result.

KONRAD WOTHE/NPL

DNA from snow tracks allowed scientists to detect the presence of the Canada lynx.

But some ecologists are reluctant to abandon conventional methods. Jean-Marc Roussel, an aquatic ecologist at the French National Institute for Agricultural Research in Rennes, says that more studies comparing the cost and accuracy of eDNA analysis to conventional monitoring methods are needed before environment-management decisions are made on the basis of eDNA results.

Molecular ecologist Cecilia Villacorta Rath at James Cook University in Townsville, Australia, thinks researchers also need to demonstrate that genetic tests are sensitive and specific enough to avoid false negatives – the failure to detect a target species that is there.

Robust results are essential because the discovery of an endangered species can have weighty legal ramifications. In the United States, such species need to be protected under the Endangered Species Act, so an area could be designated a critical habitat as a result.

As the chair of DNAqua-Net, Leese is leading the charge to develop standards that ensure genetic tests are accurate and give agencies confidence in their results. The next step could be to certify companies and laboratories doing eDNA studies, he says.

1. Day, K. *et al. Endang. Species Res.* **40**, 171–182 (2019).
2. Rees, H. C. *et al. Ecol. Evol.* **4**, 4023–4032 (2014).
3. Willerslev, E. *et al. Science* **300**, 791–795 (2003).
4. Ficetola, G. F., Miaud, C., Pompanon, F. & Taberlet, P. *Biol. Lett.* **4**, 423–425 (2008).
5. McKelvey, K. S. *et al. Fish Biol.* **88**, 1215–1222 (2016).
6. Franklin, T. W. *et al. Biol. Conserv.* **229**, 50–58 (2019).

ITALIAN PLAN FOR NEW RESEARCH AGENCY DRAWS CRITICISM

Scientists say they haven't been consulted on the creation of another national science funder.

By Marta Paterlini

The Italian government is debating whether to set up a national research agency – an organization that could boost research funding by hundreds of millions of euros a year. But although scientists have long called for such an agency, some are concerned about the latest plans. They worry that researchers haven't been involved in discussions about the organization, and that it won't be independent of political influence.

Prime Minister Giuseppe Conte, who leads a coalition government of the populist Five Star Movement and the centre-left Democratic Party, mentioned the idea for a National Research Agency (ANR) in a September speech. The proposal will be discussed in parliament this month as part of Italy's 2020 budget bill.

Italy already has several mechanisms for funding basic science, but researchers complain that the system is haphazard, and that calls for grant proposals are often delayed. The country's existing National Research Programme has a budget of €2.5 billion (US$2.8 billion) for 2015–20. But

"The agency's function and governance can only be decided after a discussion with the research community."

the scheme's main source of money for basic research — the Research Projects of National Relevance programme — last made a grant call in 2017. Moreover, Italy invests only 1.2% of its gross domestic product in research — far below the European Union target of 3%.

Many scientists had hoped for an agency that would simplify research funding, but note that the ANR instead adds another organization with its own budget. And it is not yet clear how the ANR would interact with Italy's other science-funding mechanisms. The bill up for discussion states that the agency would coordinate the direction of research at universities and public research bodies, fund "highly strategic" projects and encourage Italian participation in European and international research initiatives. It would receive €25 million in 2020, €200 million in 2021 and €300 million per year from 2022.

Missed opportunity

"It is promising that the matter is part of the current government's strategy. Unfortunately, the model behind it is not yet clear," says Vincenzo Costanzo, a cancer researcher at IFOM, a molecular-oncology institute in Milan. The move is a missed opportunity to bring all government research funding under a single body in a transparent and independent manner, he adds. "We really need an agency that regulates the annual grant calls."

Researchers also worry that they have not been involved in the ANR's planning, and are concerned about the agency's political independence. According to the bill, the ANR's leaders will be appointed mainly by politicians: the prime minister would choose the director, and government ministers would select most of the agency's eight-member executive committee. Many had instead hoped for an agency overseen by research managers and scientific advisers.

Overall, the agency is a positive step, says Giuseppe Remuzzi, director of the Mario Negri Institute for Pharmacological Research in Bergamo. But the government's role should be restricted to making suggestions about appointments, and executive-committee members should be chosen by a group operating under the best practices used by the international scientific community, he says.

Lorenzo Fioramonti, Italy's research minister, says that scientists should feed into the ANR's development. He was involved in the idea to create the agency, but says he was surprised that the draft law also included information on the agency's governance. "The agency's function and governance can only be decided after a discussion with the research community," he says. Fioramonti had hoped that the bill would serve only to set up the agency, with details of its governance and grant management decided early next year.

FIRST VACCINE AGAINST DEADLY EBOLA VIRUS WINS APPROVAL

The shot has already been given to hundreds of thousands of people in ongoing Africa outbreak.

PAMELA TULIZO/AFP/GETTY

An Ebola vaccine has been approved by the European Medicines Agency.

By Ewen Callaway

The world finally has an Ebola vaccine. On 11 November, European regulators approved a vaccine that has already helped to control deadly outbreaks of the virus — the first time any immunization against Ebola has passed this hurdle.

The decision by the European Medicines Agency (EMA) to allow US pharmaceutical company Merck to market its vaccine means that the product can now be stockpiled and, potentially, distributed more widely than it is now, particularly in Africa. In 2015, Gavi, the Vaccine Alliance — a global health partnership based in Geneva, Switzerland, that funds vaccine distribution in low-income countries — told manufacturers that it would commit to purchasing their Ebola vaccines once they had been approved by a "stringent health authority", such as the EMA.

Although several other vaccines against Ebola — a haemorrhagic fever that causes severe diarrhoea, vomiting and bleeding — are in development, Merck's is the only one that has been tested during an outbreak, in which it was shown to be highly effective at preventing infection.

The vaccine, first patented in 2003, has been administered on an emergency basis to quell the ongoing outbreak in the Democratic Republic of the Congo (DRC), which has killed some 2,000 people since it started last year. It was also used during a 2018 outbreak in that country, and in Guinea in 2015. In the current outbreak, hundreds of thousands of people have received the Merck shot, including more than 60,000 health-care workers in the DRC and several neighbouring countries.

"This is a vaccine with huge potential," said Seth Berkley, chief executive of Gavi, in a press release after the EMA's decision. "It has already been used to protect more than 250,000 people in the DRC and could well make major Ebola outbreaks a thing of the past." The organization has supported the stockpiling and delivery of Ebola vaccines and hopes to build up a global supply that could be rolled out quickly during future outbreaks.

Future protection

The EMA's approval "makes a big difference", says David Heymann, an epidemiologist at the London School of Hygiene and Tropical Medicine. But he stresses that research into the Merck vaccine and development of others must continue. "The message is that the research is not done," he adds. Research could help to develop vaccines that offer longer-lasting immunity, target more than one species of Ebola and are easier to store.

Merck's vaccine, which is marketed under the name Ervebo and known to researchers as rVSV-ZEBOV-GP, was tested in a clinical trial

conducted in Guinea towards the end of the 2014–16 Ebola outbreak in West Africa. There, the vaccine was administered to people who had been in contact with someone who was infected with Ebola, and to their subsequent contacts. It was found to offer a high level of protection against infection.

Health workers have used this strategy — known as ring vaccination — in the two other outbreaks in which rVSV-ZEBOV-GP had been deployed. But Heymann says it's important to determine whether the Merck vaccine has other uses — for instance, preventive administration to emergency health workers who might encounter Ebola in the distant future. For this, researchers will need to determine how long the vaccine's protection lasts, and whether a 'booster' dose can extend immunity.

Such studies are in the works with rVSV-ZEBOV-GP and competing vaccines, says Adrian Hill, a vaccinologist at the University of Oxford, UK. "The question remains, which vaccine would you give to, say, health-care workers to prevent them getting Ebola?"

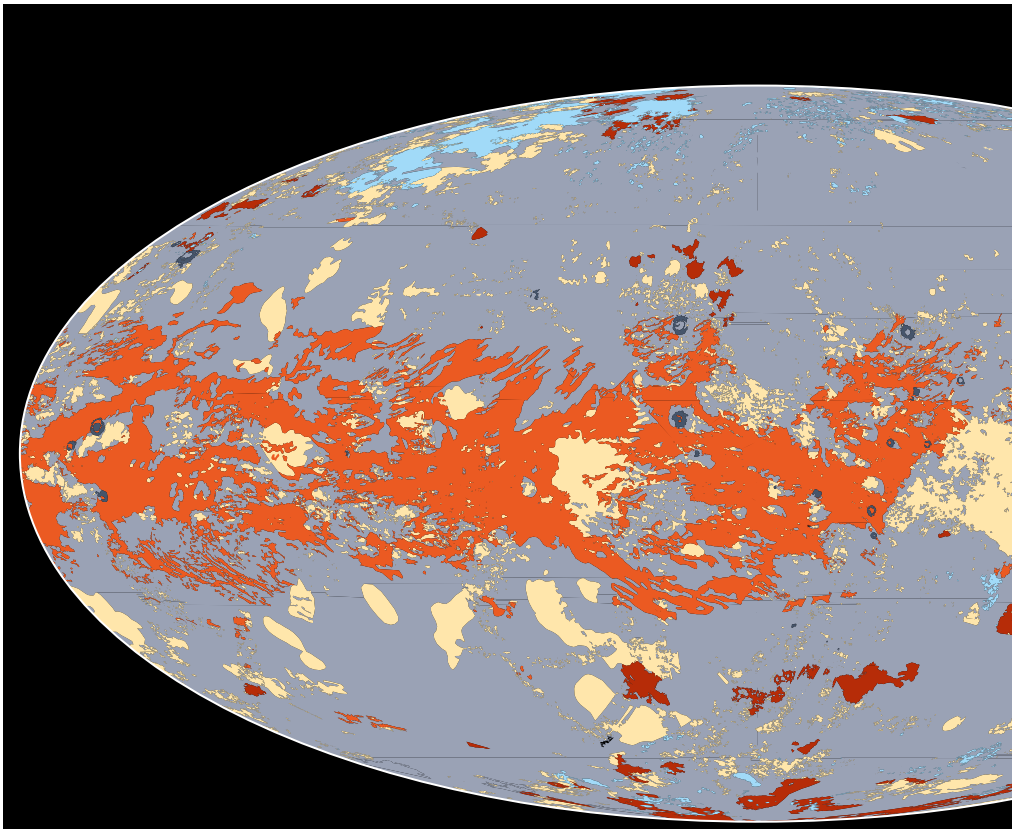
Merck's product protects against the Zaire species of the Ebola virus, which is behind the current DRC outbreak and the 2014–16 West Africa outbreak. It will be important to develop vaccines against other species of the virus — especially the Sudan species, which has caused seven known outbreaks since 1976, says Hill, who helped to test an Ebola vaccine that the London-based pharmaceutical company GlaxoSmithKline shelved in August.

There are seven other Ebola vaccines in various stages of clinical testing, according to the World Health Organization (WHO) in Geneva. In September 2019, the WHO announced that a vaccine manufactured by Johnson & Johnson in New Brunswick, New Jersey, would be used in the current DRC outbreak. Last week, the company submitted that vaccine for EMA approval.

Unlike the Merck vaccine, which is given in one dose, the Johnson & Johnson immunization requires a booster shot that is administered 56 days after the first injection. In the DRC, it will be given to people at risk of Ebola, such as health-care workers, in areas where the virus is not already circulating.

And next month, Gavi's board will decide whether to establish a global stockpile of Ebola vaccines. Merck, which is headquartered in Kenilworth, New Jersey, is seeking approval for its vaccine by the US Food and Drug Administration.

On 12 November, the WHO announced it had "prequalified" the Merck vaccine, which means that the product meets the agency's standards for quality, safety and efficacy. Other UN agencies, Gavi and many national health agencies look to this endorsement when procuring and delivering a vaccine.



SOURCE: R. M. C. LOPES *ET AL. NATURE ASTRON.* HTTP://DOI.ORG/DFB8 (2019)

A map of Saturn's largest moon

Astronomers have used data from NASA's Cassini mission to map the entire surface of Titan, Saturn's largest moon, for the first time. Their charts reveal a diverse terrain of mountains, plains, valleys, craters and lakes unlike anywhere in the Solar System outside Earth.

The Cassini spacecraft orbited Saturn from 2004 to 2017 and collected vast amounts of information about the gas giant and its moons. The mission included more than 100 fly-bys of Titan, allowing researchers to glimpse the moon's surface through its thick atmosphere and survey its terrain in unprecedented detail.

Rosaly Lopes, a planetary scientist at NASA's Jet Propulsion Laboratory in Pasadena, California, and her colleagues stitched together images and radar measurements taken by the spacecraft to produce the global map of Titan, which they published on 18 November in *Nature Astronomy* (R. M. C. Lopes *et al. Nature Astron.* http://doi.org/dfb8; 2019).

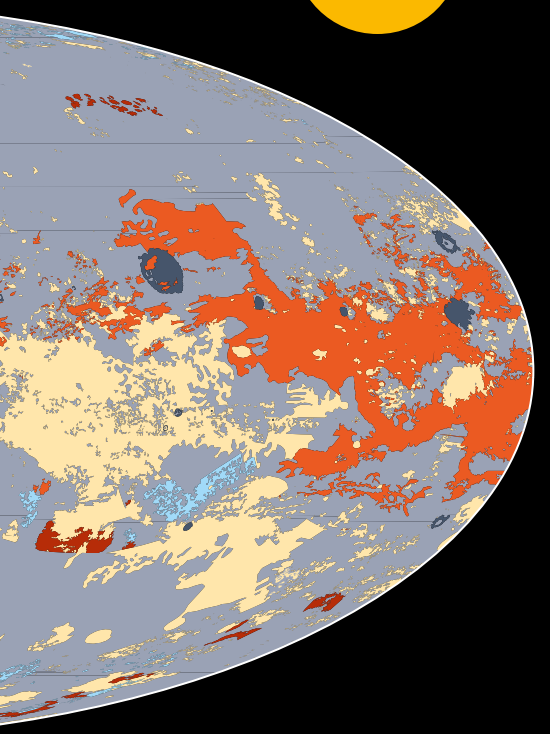
"Titan has an atmosphere like Earth. It has wind, it has rain, it has mountains. It's a really very interesting world, and one of the best places in the Solar System to look for life," says Lopes.

Nearly two-thirds of Titan's surface consists of plains, the map reveals, and 17% is covered in sandy dunes shaped

Geographical features

- Lakes
- Craters
- Dunes
- Hummocky
- Labyrinth
- Plains

Data story



by the wind, mostly around the equator. Around 14% of the surface is classified as 'hummocky' — hilly or mountainous — and 1.5% is 'labyrinth' terrain, with valleys carved by rain and erosion. There are surprisingly few impact craters, suggesting that the moon's surface is fairly young.

Titan is the only world in the Solar System aside from Earth with known bodies of liquid on its surface. However, these seas and lakes are filled with liquid methane rather than water, and they cover just 1.5% of the moon's surface.

"The most profound discovery of Cassini is that Titan is so diverse," says Ralph Lorenz, a planetary scientist at the Johns Hopkins University Applied Physics Laboratory in Laurel, Maryland. "It's almost like a completely different world."

By 2034, NASA plans to send a drone to Titan on the Dragonfly mission, which will fly across the surface and study it in multiple locations. But there are no current plans to send further orbiters to Saturn or its moons, so this map is likely to remain our best global view of Titan for the foreseeable future.

By Jonathan O'Callaghan

LAB SEQUENCES GENOMES OF A CONTINENT'S BUTTERFLIES

Draft genomes of more than 800 varieties hint at the role of interbreeding in the animal's evolution.

By Ewen Callaway

When biologist Nick Grishin wanted to tackle big questions in evolution — why some branches of the tree of life are so diverse, for instance — his team set out to sequence the genomes of as many butterflies as it could: 845 species, to be precise.

In a study that some researchers are hailing as a landmark in genomics, Grishin's group at the University of Texas Southwestern Medical Center in Dallas sequenced and analysed the genome of what it called a "complete butterfly continent": every species of the creature in the United States and Canada. The study was posted on the bioRxiv server on 4 November[1].

"I think its bloody amazing, because the technology involved in sequencing 845 species is there," says James Mallet, an evolutionary biologist at Harvard University in Cambridge, Massachusetts. "It's a beautiful piece of work."

The data allowed Grishin's team to build an evolutionary tree detailing the relationships of all the butterflies, as well as to determine the pace at which new species formed. The team suggests that fast-diversifying groups of butterflies are those that swap genes with close relatives through interbreeding — a phenomenon that could extend to other organisms.

Others, however, point out that most of these genomes will be of limited use to other researchers, because they are low-quality 'drafts' comprised of thousands of short DNA stretches, and not higher-quality sequences that have been assembled into longer stretches. Grishin says that the sheer number of genomes, even of low quality, allows his team to draw broad conclusions about evolution that could not be made from more limited data sets. He plans to make the genomes publicly available.

Butterfly patterns

Grishin, whose research group studies the shape and evolution of proteins, started researching butterflies after reading a 2012 paper[2] on the diverse tropical genus *Heliconius*, whose species have elaborate wing patterns that mimic those of other butterflies. The study found that some genes that determine wing patterns seemed to have

TIM ZUROWSKI/SHUTTERSTOCK

A *Heliconius* butterfly.

been passed between three *Heliconius* species through interbreeding, instead of being inherited from the species' common ancestor, and suggested that such swaps explain the huge diversity of *Heliconius* butterflies.

Inspired by that work, Grishin wondered whether such a connection could be seen in other butterflies. "Some groups diversify very rapidly and there are many species in them, and others are kind of empty," he says. "So to understand why and how that happens, we would need to sequence them all."

At one time, sequencing hundreds of butterfly genomes would have been unaffordable, but costs have plummeted in recent years. Collecting samples for every species in the United States and Canada was still a challenge, however. Grishin's team worked with amateur butterfly enthusiasts as well as museum collections across the United States to gather data – a single leg from a dead specimen was enough to obtain a draft-quality genome.

Once they had sequenced the genomes of all 845 species, the researchers worked out the evolutionary relationships. Their butterfly family tree broadly agreed with existing ones based on anatomy and more limited genetic analyses, although the group did reclassify 40 species and suggested several new groupings at the genus level.

The tree also revealed that some butterfly groups have evolved faster than others. Two of the fastest-evolving ones, commonly known as the blues and the whites, have developed highly specialized interactions with other organisms that might explain their rapid evolution, say Grishin's team. The blues, or Polyommatinae, form symbiotic relationships with ants, whereas whites, or Pierini, have developed adaptations to feed on mustard plants that are toxic to many other insects.

"Some groups diversify very rapidly. To understand how that happens, we would need to sequence them all."

An analysis of genes shared by multiple species also showed that these diverse groups were likely to have acquired genes through interbreeding. Many of the genes that are swapped between species are thought to be involved in mate recognition and other factors that can cause species splits. Grishin says that by spreading such genes, interbreeding – rather than the gradual accrual of new mutations – could be helping to drive the evolution of butterfly species.

The link between interbreeding and speciation is "an idea that is sort of coming to the fore", says Mallet, who co-led a team that reported similar findings in *Heliconius* butterflies this month[3].

Missing data

In addition to the draft genomes, Grishin's team generated 'reference' genomes, in which genes are assembled into chromosome sequences, for 23 species.

High-quality genomes such as this are the targets of other large-scale projects to sequence the tree of life. In 2018, a consortium called the Earth BioGenome Project laid out plans to decode the genomes of the roughly 1.5 million known species of eukaryote – animals, plants, protozoans and fungi – at an estimated cost of US$4.7 billion over 10 years.

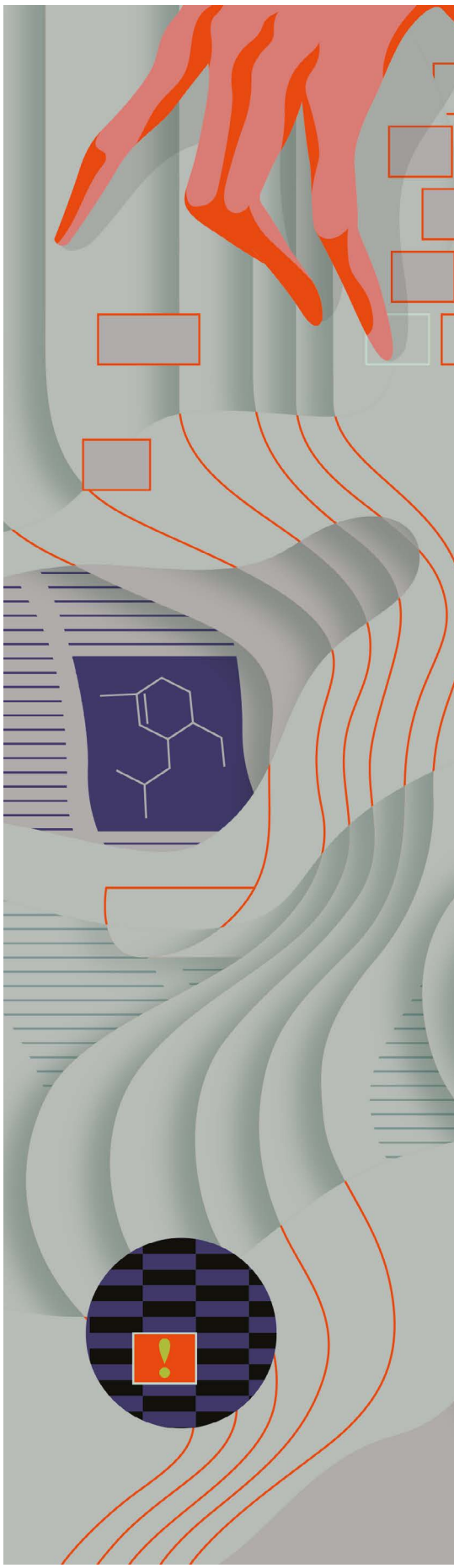
Grishin is enthusiastic about these efforts, particularly for vertebrates. But he thinks there are too many unknown species of invertebrate to sequence them all in the near future. "I don't think they will succeed very quickly," Grishin says. "Our efforts – where we just jump right in and do things right away without much fuss about it – may be helpful."

1. Zhang, J., Cong, Q., Shen, J., Opler, P. A. & Grishin, N. V. Preprint at BioRxiv https://doi.org/10.1101/829887 (2019).
2. Heliconius Genome Consortium. *Nature* **487**, 94–98 (2012).
3. Edelman, N. B. *et al*. *Science* **366**, 594-599 (2019).

THE INTEGRITY INSPECTORS

Some research institutes now pay for independent screening of their scientists' manuscripts.

By Alison Abbott



On 15 June 2017, scientists at a respected biological institute in Germany were thrown into crisis by an alarming announcement. An investigation into the Leibniz Institute on Aging had found that its director, cell biologist Karl Lenhard Rudolph, had published eight papers with data errors, including improperly edited or duplicated parts of images.

Investigators didn't find deliberate fraud, but Rudolph wasn't able to present original data to explain the problems. The Leibniz Association, which runs the institute in Jena and had commissioned the probe, concluded that Rudolph hadn't supervised his lab group properly, and so was guilty of "grossly negligent scientific misconduct". It applied the strictest sanctions it could, barring the institute from applying for research funding from the association while under Rudolph's leadership for three years. It also ordered the centre to undergo an international review, even though the last one had been completed only a couple of years earlier. Rudolph resigned as director.

It was the second calamity in a year for the centre, which is also known as the Fritz Lipmann Institute (FLI). Police had raided it in 2016 after allegations that the centre had violated European regulations on animal experiments. The experiments were suspended, and

ILLUSTRATION BY ANA KOVA

although the FLI was cleared of the allegations, not all of the experiments had been re-authorized when the Rudolph affair broke. "The second crisis sent us into shock – it seemed more personal," says molecular geneticist Christoph Englert, a group leader at the FLI, which employs 270 scientists. Most researchers at the centre hadn't even known their director was under investigation.

FLI leaders set about restoring the centre's reputation. They began by phasing in mandatory electronic databases and creating a system of thesis advisory committees to replace single PhD supervisors. The FLI's head of core facilities, Matthias Görlach, had a less conventional idea. He contacted Enrico Bucci, a molecular biologist who had visited the FLI for some PhD work 18 years earlier, and with whom he'd kept in touch. Bucci was now in the business of checking research papers, Görlach knew; in 2016, he'd founded a science-integrity firm called Resis, based in Samone, Italy. Could the company perhaps help the institute to avoid errors in future?

So began a remarkable system of outside vetting, in which researchers at the FLI must send every paper and master's thesis across to Resis for screening before they submit them for publication. It's an unusual step. Some journals check papers for errant statistics or manipulated images before publishing, but most research institutions say it's up to the scientists themselves to ensure their manuscripts are correct. "I am not aware of any US institute doing this," says Lauran Qualkenbush, director of research integrity at Northwestern University in Chicago, Illinois, and president of the US Association of Research Integrity Officers.

And some researchers disapprove. "The moment an institution needs to constantly question the moral integrity of its scientists by double-checking submitted figures, the leadership should resign," says Giulio Superti-Furga, director of the Research Center for Molecular Medicine in Vienna.

But amid rising concern about the quality and reproducibility of research, particularly in the biomedical sciences, a handful of European institutions have told *Nature* that they have now hired external companies or dedicated in-house experts to check research manuscripts. The institutions say the cost of the endeavour is worthwhile, not only for

> **"NOW THAT WE SUBMIT TO EXTERNAL CHECKING, I HAVE MORE CONFIDENCE."**

the immediate benefit of the checks, but also because it can help them to spot areas in which their scientists need extra training.

Scientists at the FLI and other institutions see the extra layer of checks as protective, not intrusive. "Because of the manuscript check, I sleep at night," says one FLI group leader, Björn von Eyss. "I had started to worry about whether I had done something wrong in my papers, maybe missed a label: a mistake can become misconduct," adds Lilia Espada, a postdoctoral researcher at the centre. "Now that we submit to external checking, I have more confidence."

Science under scrutiny

Across the research world, there is growing suspicion about sloppiness and outright misconduct in the scientific literature. The number of retractions of research papers has risen to around 1,400 a year, compared with about 40 at the turn of the millennium, notes Ivan Oransky, a journalist in New York City who co-founded the website Retraction Watch, which monitors and reports on retractions.

In 2016, Elisabeth Bik, a microbiologist then at Stanford University in California, reported that around 4% of more than 20,000 biomedical papers she had examined contained inappropriately duplicated images. (Bik is now a full-time research-integrity consultant.) And last year, Bucci reported that about 6% of a sample of 1,364 papers he had looked at contained at least one instance of image manipulation.

Increasingly, fraud-busters are starting to hunt down manipulated images in published papers and flag them widely. Rudolph's work is an example: the faults were exposed by an external whistle-blower, who sent the findings to Rudolph, then to the DFG, Germany's main national funding agency, and to its independent Ombudsman Commission. The Leibniz Association has declared a zero-tolerance approach, and young scientists at the FLI say they feel under pressure. Some have told *Nature* privately that they are worried because of the way in which even unintended errors in papers are flagged publicly online. It can be easy to make a mistake when handling massive and complex biological data sets, they say — and they fear their papers might be publicly picked apart, derailing their careers before they get started.

In this atmosphere, the idea of the sort of pre-submission screen that Bucci's company was offering appealed to Görlach. Bucci had been drawn into the world of research integrity after founding an image-search company called BioDigitalValley in Pont-Saint-Martin, Italy, in 2008 that aimed to sell a service to biomedical scientists who wanted all images relevant to a particular tissue or disease extracted from the literature. Bucci had first made a giant database of accessible biomedical papers and cleared it of retracted articles. He then checked the images in all publications by the authors of those retracted papers. He found serious problems in the work of many of them, particularly that of Alfredo Fusco, a then-prominent cancer researcher at the University of Naples Federico II. Fusco has now had 24 papers retracted and 10 corrected. The affair, which implicated scientists in Fusco's network at other institutes in Italy and beyond, sent shock waves through the scientific community. Bucci was so disturbed by what he saw that he switched career path, founding Resis, to try to do something about it.

Restoring reputation

After Görlach contacted him, Bucci gave the FLI's group leaders a presentation of his work. His company's proprietary software scans images in a manuscript for duplication or unlikely composition, he told them. Resis has just two employees, but brings in consultants for particular contracts. In late 2017, FLI group leaders sent Bucci some sample papers and theses to check — and were impressed by the results. He picked up some small errors they hadn't spotted. The institute signed a contract with Resis to analyse the images in all papers, to do random checks on statistics and also, in master's theses, to look for plagiarism. Resis screens all manuscripts within 24 hours of receipt, although if the screen flags problems, further analysis can take up to three more days. The institute budgets up to €50,000 (US$55,000) per year to cover both the service and its handling of the information that Resis supplies.

> "IT IS VERY IMPORTANT FOR OUR IMAGE AS AN INSTITUTE TO GET BACK ON TRACK."

The new system began in April 2018, and the first results proved its value, says molecular geneticist Alfred Nordheim at Germany's University of Tübingen, who became the FLI's interim scientific director when Rudolph stepped down. Resis found no serious problems in the first 40 manuscripts that it analysed for the institute, but it did flag at least one issue in 17 of them, Nordheim says. "Most of these issues were to do with the use of statistics — things like undersampling or use of not fully suited statistical procedures," he says. "The Resis analysis has been important for us because it allowed us to identify patterns of errors, and act accordingly." Now, for example, the institute has introduced mandatory statistical workshops for all of its scientists.

FLI researchers see the system as a positive step that is helping to protect them from error. Rudolph himself says that had the checking system had been in place earlier, he would have caught the problems in his papers. (Five have been corrected, one remains under discussion at a journal, and in two cases, journal editors decided no correction was needed, he says.) Rudolph remains a lab leader at the FLI, but his group has now shrunk to seven scientists, half the size it was before the scandal broke.

In June this year, Marco Foiani, the scientific director of IFOM, a molecular oncology research institute in Milan, Italy, learnt about the initiative during a meeting of the FLI's international scientific advisory board, of which he is a member. It struck an immediate chord with him: IFOM was itself reeling from research misconduct investigations involving a former director, Pier Paolo Di Fiore, who had co-authored some papers with Fusco that have been retracted. Di Fiore says he agrees with the retractions, but wasn't involved in putting the figures together for the papers. IFOM had introduced electronic notebooks and other measures to promote good scientific practice, and Foiani decided to add on external checking, also using Resis. "It is very important for our image as an institute to get back on track," says Foiani.

As at the FLI, young researchers at IFOM welcome the screens. "Having a research scandal can affect the credibility of the whole institute," says Ylli Doksani, one of IFOM's 24 research group leaders. "We are mostly funded by a charity, and I am happy if the institute does whatever is needed to maintain trust and show we take integrity issues very seriously."

Other organizations have decided to do publication checks internally. After the Beatson Institute in Glasgow, UK, had to deal with a retraction in 2012, it hired a dedicated integrity offer, former molecular biologist Catherine Winchester, to check all papers destined for publication by eye. "It took only a short time for the more junior scientists to shed their fear that they were being policed, but there was immediate buy-in from senior PIs," she says. "Now everyone is really grateful for the service."

The cost of checks

Some research organizations rule out external checks for themselves. The president of Germany's Max Planck Society, Martin Stratmann, says that the society — which runs 78 elite research institutes — does not need to commission outside checkers because research directors themselves have the mandate and responsibility to check every paper before it goes out. Some institutes *Nature* talked to for this story were unwilling to comment on the topic; others said only that they found it interesting. "We will monitor

FLI/ANNA SCHROLL

Matthias Görlach helped to set up a manuscript vetting system at the FLI in Jena, Germany.

the process and discuss with our faculty," says Bruce Stillmann, director of the Cold Spring Harbor Laboratory in New York.

Nor do all institutes that have been hit by a scandal see the need for screening. In 2012, the DFG judged that Silvia Bulfone-Paus, a senior scientist at another Leibniz institute, the Research Center Borstel, had failed in her supervisory duties after data manipulation was discovered in more than a dozen of her papers. Centre director Stefan Ehlers doesn't think that paying for independent checks is the right way to approach these problems: rather, he says, it's important to foster "a culture of trust and fearlessness to report mistakes and to discuss questionable data".

And pre-submission checks wouldn't stop all types of fraud, adds Shinya Yamanaka, a Nobel laureate and director of an institute that has recently experienced such a case, the Center for iPS Cell Research and Application at Kyoto University in Japan. There, in 2018, stem-cell researcher Kohei Yamamizu was found guilty of fabricating and falsifying images in a high-profile paper in *Stem Cell Reports*. Yamanaka implemented measures such as electronic notebooks and mandatory storing of all experimental data – but did not opt for pre-submission checks, a method that "does not investigate whether experiments were truly carried out and recorded appropriately", he told *Nature* in an e-mail.

Still other institutions say that the checks are beyond their budget. The Italian National Research Council (CNR), which runs 102 research institutes, would like to offer a full – but voluntary – screening service to its institutes, but says it can't afford to. After the Fusco affair, it established a technical unit to use licenced Resis image-analysis software to check published papers. The unit provided formal comments on the report of the University of Naples' investigation into Fusco's papers, and now focuses on allegations of misconduct by CNR researchers. If an allegation surfaces, the unit examines all the papers the institute in question has published over the previous five years. Any manipulated images are recorded in a growing database.

Last year, the CNR unit started preventive work on a modest scale: it has so far done a few pre-submission checks, responding to individual CNR researchers who were concerned, for example, about joining in as co-authors on particular manuscripts. "Prevention is the critical step," says Cinzia Caporale, who leads the organization's research integrity activities from its headquarters in Rome. After the scandals in Italy, "scientists don't always trust their colleagues any more", she says. Caporale thinks the CNR's work has increased scientists' awareness: the council's database suggests that its scientists are already publishing fewer inappropriate images, she says. A higher budget would allow more systematic pre-checking, but Caporale says there is no prospect of that right now.

Not many image-checking services have the capacity to rapidly screen a high volume of papers, as an institute – or a journal – might require. But some say they are interested. Sheridan, a large publishing-services firm in Hunt Valley, Maryland, already offers image forensics to journals, and told *Nature* that it is "open to the idea" of setting up such a service for institutions. Mike Rossner, who runs a small consultancy firm called Image Data Integrity in San Francisco, California, says he'd prefer to train someone from an institution's research integrity office to do screening using his own manual system. Rossner is known for his expertise in spotting problems in papers by eye: as a former managing editor at the *Journal of Cell Biology*, he introduced checks of images in all papers accepted for publication – making the journal the first major life-sciences publication to institute the practice.

Nurturing trust?

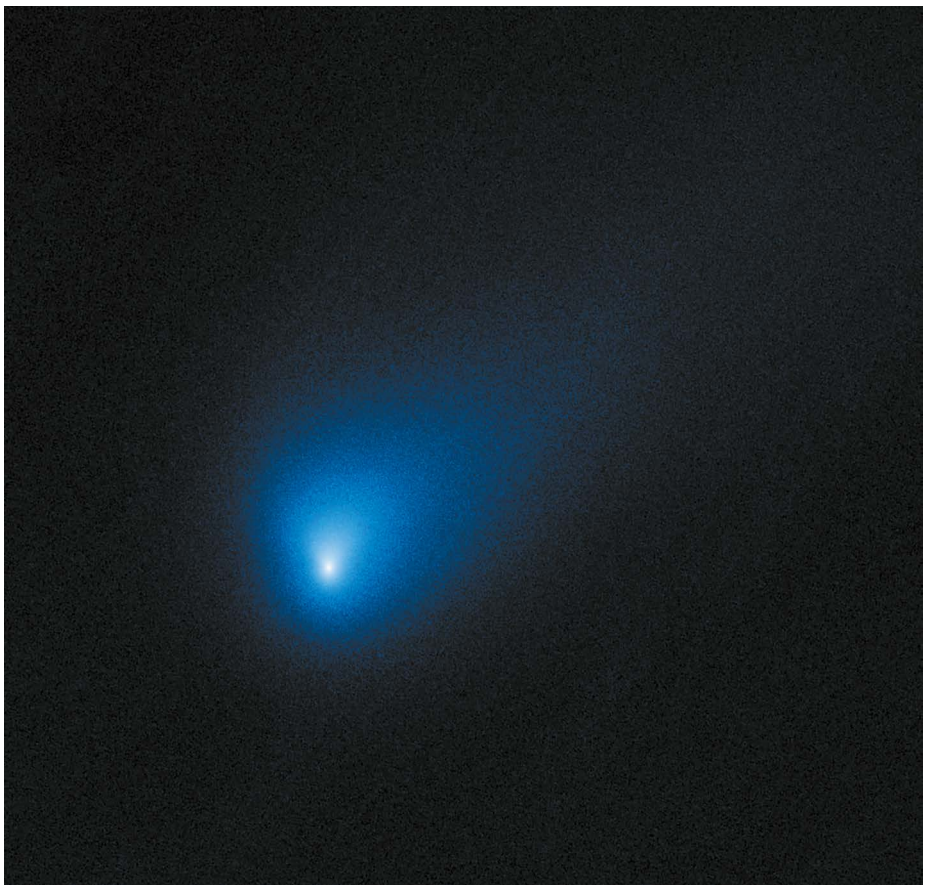
Rossner thinks that investing in pre-checking could save money in the long run. "Prophylactic screening makes financial sense, because any case brought against an institution for publishing misleading data could cost an institute even more in legal fees," he says. It might even become a selling-point for institutes, suggests Caporale. "Being able, for instance, to tell journal editors that a paper has been independently checked may nurture trust," she says.

Even if that were true, it wouldn't relieve journals of the responsibility to do their own checking, says Bernd Pulverer, chief editor of the *EMBO Journal* in Heidelberg, Germany. His journal checks images in all papers before they are accepted, and generally sees problems in around one in five manuscripts, a proportion that has not changed since the journal began the checks ten years ago, he says. Only a tiny minority (0.5%) of these involve outright fraud. Other journals now regularly check images too, although some (including *Nature*) do spot checks, not systematic ones.

Journals don't have the same jurisdiction as a scientist's employer does to investigate problems, so the institute has an important role in ensuring quality, Pulverer adds. "But it is important for the employer not to start over-policing, because that can backfire," he says.

The FLI plans to continue working with Resis and thinks that the checks will make the institute more attractive in competing for the best scientists, says Nordheim. In June 2018, it reported its experience to a Leibniz Association leadership meeting on good scientific practice. Matthias Kleiner, the association's president, was impressed. He is planning to test the possibility of introducing a certification system for good scientific practice for the association's institutes. It's possible that pre-submission checks could be an optional item on these certificates. For some Leibniz institutes, he adds, it could be a way "to protect scientists from being in danger of scientific misconduct".

Alison Abbott writes from Munich, Germany.



A composite image of interstellar Comet 2I/Borisov, taken by the Hubble Space Telescope.

NASA/ESA/D. JEWITT (UCLA)

INTERSTELLAR INTRUDERS

Astronomers grapple with the meaning of the first two objects entering our Solar System from distant regions. **By Alexandra Witze**

From the tallest peak in Hawaii to a high plateau in the Andes, some of the biggest telescopes on Earth will point towards a faint smudge of light over the next few weeks. The same patch of sky will draw the attention of Gennady Borisov, an amateur astronomer in Crimea, and many other hobbyists who will sacrifice proper sleep and doze through their day jobs rather than miss this golden opportunity.

What they're looking for is a rare visitor that is about to make its closest approach to the Sun. After that, they have just months to grab as much information as they can from the object before it disappears forever into the blackness of space.

This chunk of rock and ice started its journey many light years from Earth, millions of years ago. The object got kicked out of its own neighbourhood by a violent gravitational push — maybe from a nearby planet, maybe from a passing star. Since then, it has been adrift in the space between the stars, eventually heading in our direction.

On 30 August, Borisov spotted the object in the predawn sky — it was glowing dimly, with a broad stubby tail. Later named Comet 2I/Borisov after its discoverer, it captured global attention because it's only the second object — aside from exotic dust particles — ever known to have entered our Solar System from interstellar space. "This is my eighth comet, and so amazing," says Borisov, who adds that it was "great luck that I got such a unique object".

It is remarkably different from the first interstellar interloper, which was a small, dark, rocky-looking object named 1I/'Oumuamua that whizzed past the Sun in 2017. Together, these two interstellar objects are rewriting what researchers know about the icy bodies — estimated to number as many as 10^{26} — that float unmoored throughout the Milky Way.

Among other things, 1I/'Oumuamua and 2I/Borisov have provided the first direct glimpse of the physics and chemistry of the squashed debris clouds that surround young stars and serve as the birthing grounds for planets. These samples from other planetary systems are allowing scientists to explore whether the Solar System is unique or whether it shares building blocks with other planetary systems in the Milky Way.

Because astronomers spotted 2I/Borisov on its way into the Solar System, they have many months to study it — unlike their fleeting glimpse of 'Oumuamua, which was discovered on its way out. As a result, they expect to learn much more from 2I/Borisov, such as what chemical compounds make up its icy heart. It is their best look yet at an object known to have formed around another star.

And as telescopes continue to probe the sky for faint, fast-moving objects, researchers expect that they will spot many more interstellar interlopers in coming years. "It's been so much fun to see this suddenly crack open and watch a new field develop," says Michele Bannister, a planetary astronomer at Queen's University Belfast, UK.

Dusty origins

Interstellar objects probably began their lives when icy grains clumped together in a disk of gas and dust around a young star. These are the same regions where planets grow from small nuclei and then ping-pong into different orbits around the star because of collisions and gravitational shoves.

The planets push through the icy rubble like a snowplough shouldering its way through a pile of hailstones. And modelling results suggest that the planets fling more than 90% of those 'hailstones' out of their star's sphere of influence and into interstellar space. There they drift, as lonely scattered objects, until they happen to pass close enough to another star

to be attracted by its gravity for a quick visit.

Astronomers had expected that the first interstellar object they saw would look like a typical comet. Most comets in the Solar System hail from the distant realm known as the Oort cloud, a sort of cosmic deep freeze that lies roughly 1,000 times farther away from the Sun than Pluto. Occasionally, something perturbs one of these comets and sends it careering towards the Sun; as it gets closer and warms up, its nucleus sprays out dust and gas that form a classic cometary tail.

But when the first interstellar visitor showed up, it didn't look like a conventional comet. Unlike them, 'Oumuamua was tiny – just 200 metres or so across – and rocky. Also, it was shaped like a cigar and tumbling end over end. That's about all scientists could work out before 'Oumuamua headed out of the Solar System[1].

By contrast, 2I/Borisov looks like an ordinary comet – and researchers are taking advantage of their time to study it (see 'Dropping by'). "We are keenly interested in seeing what the chemistry of this comet is, to see if it is different from those in the Solar System," says Karen Meech, an astrobiologist at the University of Hawaii in Honolulu.

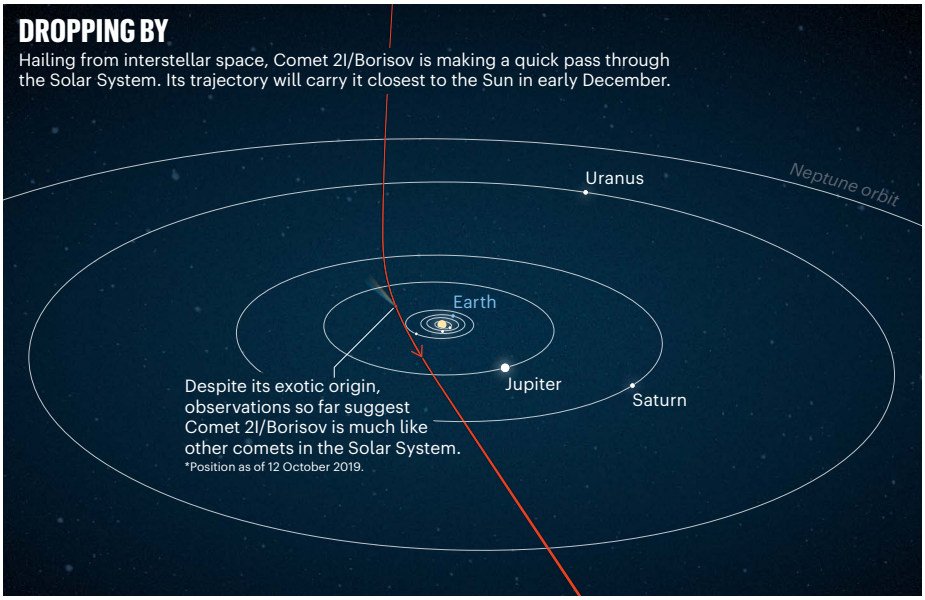
2I/Borisov is reddish in colour and is steadily spraying out dust particles[2,3]. Its nucleus is relatively small, perhaps just one kilometre across, but that's not unheard of for Solar System comets.

"After 'Oumuamua, we had to completely revise what we thought interstellar objects might be like," says Matthew Knight, a comet specialist at the University of Maryland in College Park. "But now the second one coming through looks more or less, so far, like what we thought we might see from a comet ejected from another star. Now I feel a lot better." That suggests that the star systems where other worlds form might be much like our own.

The discoveries are coming fast. Just three weeks after 2I/Borisov was first seen, astronomers trained the 4.2-metre William Herschel Telescope in Spain's Canary Islands on it and spotted molecules of cyanide gas streaming off the comet[4]. It was the first-ever detection of gas from an alien visitor to the Solar System.

On 11 October, another research team used a 3.5-metre telescope in New Mexico to detect oxygen coming off the comet[5]. The oxygen probably came from water breaking apart in the comet's nucleus, making this the first time that researchers have spotted water from another star system entering our own. Together, the amounts of cyanide and water spraying from the comet aren't surprising compared with what astronomers have seen from many other bodies.

Astronomers are watching keenly to see what other molecules, such as carbon monoxide, they can spot coming off 2I/Borisov as it gets closer to the Sun and warms up, which will further reveal how similar – or how different – it is to comets in the Solar System, says Maria Womack, an astronomer at the Florida Space Institute at the University of Central Florida in Orlando.



ADAPTED FROM ILLUSTRATION BY NASA/ESA/J. OLMSTED AND F. SUMMERS (STSCI)

Early observations also suggest that 2I/Borisov might contain relatively low amounts of carbon-chain molecules such as C_2 and C_3 (ref. 6). About 30% of the comets in the Solar System are similarly carbon-depleted. They typically come from relatively close to the Sun, rather than from the far reaches of the Oort cloud.

As months pass and astronomers gather more observations of 2I/Borisov, they hope to be able to understand much more about the planet-forming disk where it originated. "It's going to be really exciting to figure out what the building blocks of other systems are going to look like relative to ours," says Malena Rice, a graduate student in astronomy at Yale University in New Haven, Connecticut.

"They're like buses. You wait decades for one to come along, and then two come along almost at once."

Researchers also hope to start unravelling how interstellar objects might have voyaged through deep space before showing up in the Solar System. Estimates suggest the objects experience many forces as they orbit the centre of the Galaxy, including occasional encounters with other stars or nudges from Galactic tides. Some scientists have tried to calculate which stars 1I/'Oumuamua and 2I/Borisov could have formed around, but tracing their orbits back is difficult[7] – like trying to reconstruct which bar a London pub-crawler started at from the final one they visited.

Other questions include when we can expect the next interstellar visitor, and how different it might be from 1I/'Oumuamua and 2I/Borisov. Scientists didn't expect two in such rapid succession after decades of fruitless searching. "I remain confused and astounded that the second object came along so fast," says Robert Jedicke, an asteroid specialist at the University of Hawaii, who has worked to calculate the frequency of interstellar visitors[8]. "They're like buses," says Alan Fitzsimmons, an astronomer at Queen's University Belfast. "You wait decades for one to come along, and then two come along almost at once."

Some astronomers are now poring through archival data to see whether objects spotted years ago were actually interstellar visitors that researchers did not recognize at the time. And the future rate of discovery is expected to rise – perhaps to one interstellar object a year – when the Large Synoptic Survey Telescope goes online in Chile in 2022, from where it will survey the entire visible sky every three nights. The European Space Agency has been working on a spacecraft concept, known as Comet Interceptor, that could visit future interstellar objects as they wing their way past the Sun.

Once astronomers have 10 or 20 interstellar objects under their belts, they should have a much better picture of what these deep-space wanderers are really like. "Eventually we'll be talking about the Galaxy as something in which we are exchanging the products of planetary systems," says Bannister. "It will be an entirely different way of doing astronomy."

Alexandra Witze writes for *Nature* from Boulder, Colorado.

1. Bannister, M. T. *et al. Nature Astron.* **3**, 594–602 (2019).
2. Bolin, B. T. *et al.* Preprint at http://arxiv.org/abs/1910.14004 (2019).
3. Guzik, P. *et al. Nature Astron.* https://doi.org/10.1038/s41550-019-0931-8 (2019).
4. Fitzsimmons, A. *et al. Astrophys. J. Lett.* **885**, L9 (2019).
5. McKay, A. J., Cochran, A. L., Dello Russo, N. & DiSanti, M. A. Preprint at https://arxiv.org/abs/1910.12785 (2019).
6. Opitom, C. *et al.* Preprint at https://arxiv.org/abs/1910.09078 (2019).
7. Hallatt, T. & Wiegert, P. Preprint at https://arxiv.org/abs/1911.02473 (2019).
8. Engelhardt, T. *et al. Astron. J.* **153**, 133 (2017).

Books & arts

Far sight and flotsam

Two books explore how humans are both probing and polluting outer space. **By Meg Urry**

Dr Space Junk vs The Universe: Archaeology and the Future
Alice Gorman
MIT Press (2019)

Handprints on Hubble: An Astronaut's Story of Invention
Kathryn D. Sullivan MIT Press (2019)

The space age erupted with a flurry of satellites. The first two Soviet Sputniks launched in 1957, soon followed by the US Explorer I and Vanguard I. In 1959, spurred on by cold-war tensions, NASA selected seven men as astronauts for its Project Mercury programme. (Thirteen women who passed the same hurdles, courtesy of a private, parallel programme, were vetoed.) Barely a decade later, NASA's Apollo astronauts walked on the Moon.

The dawn of space exploration was all about the now. It was daring and risky, punctuated by engineering miracles and an air of invincibility. It wasn't, however, focused on the long term. Now, six decades on from Sputnik, old spacecraft are displayed in museums, robotic missions regularly reveal secrets from throughout the Universe and private companies such as SpaceX are planning colonies on Mars. And the rich, crowded future of space is the focus of two books, one by space scientist and oceanographer Kathryn Sullivan, the other by space archaeologist Alice Gorman. Both make us think more deeply about how we, as humans, ought to fit into the cosmos.

In *Handprints on Hubble*, Sullivan, a former NASA astronaut who helped to launch the Hubble Space Telescope in 1990 and has been involved in updating its capabilities since, highlights the importance of planning for new instruments and infrastructure. Gorman, meanwhile, applies an archaeologist's perspective to space-related materials and activities in *Dr Space Junk vs the Universe*.

Hubble has made more than one million observations of stars and galaxies, and probed dark matter and the history of the Universe itself, over nearly three decades. I worked for many years at the Space Telescope Science Institute in Baltimore, Maryland, which runs Hubble's science operations for NASA. I was there for the launch, the discovery of a flaw in the primary mirror and the astonishing fixes that astronauts repeatedly pulled off. Yet Sullivan's book makes clear how much I hadn't known. Hers is a first-hand story, from conception to today, of the first space mission for which in-orbit maintenance and repair were integral from the start.

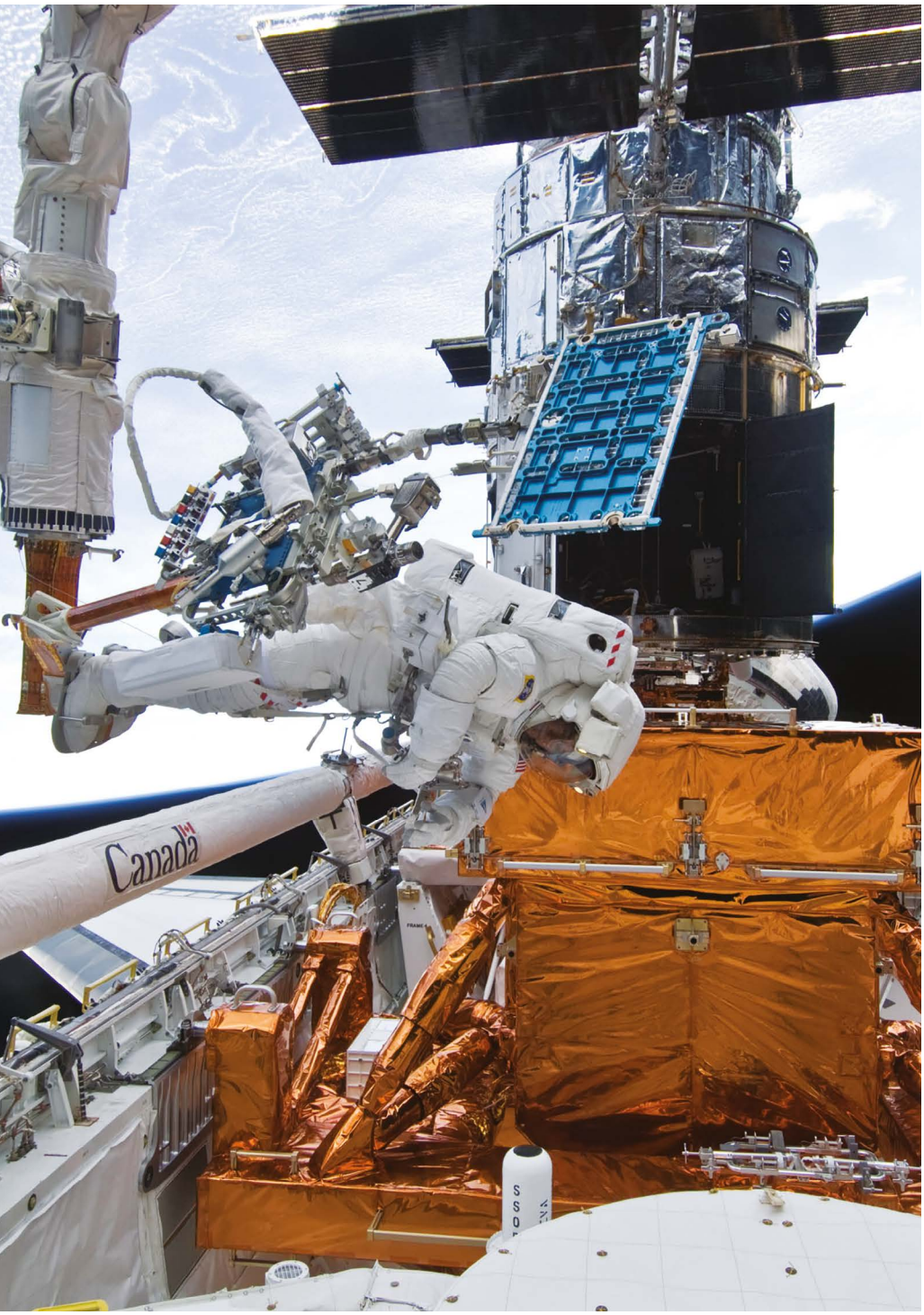
Sullivan brings alive the strenuous challenges of space mechanics. Replacing entire instruments or — much harder — parts deep inside them during long, arduous spacewalks demands custom-designed tools. For example, Sullivan explains the evolution of the foot anchors that keep astronauts in place. Without these, turning a screw one way would make the astronaut and/or the spacecraft rotate in the opposite direction. This is the kind of detail that underscores the complexity of the job.

Every step needs forethought. Once removed, a screw will float away if not caught, creating dangerous space junk that could damage other craft, as Gorman discusses. Sullivan and her colleagues spent hundreds of hours testing tools and procedures on a simulated Hubble in an underwater tank, with scuba-diving gear standing in for unwieldy spacesuits.

The meticulously planned servicing missions are what have kept Hubble at the leading edge. Its first set of instruments was selected in 1978. By today's standards, the technology was impossibly crude and the computer storage limited. After the first servicing mission in 1993, new optics compensated for the flaw in the mirror, the workhorse camera had improved detectors and the spacecraft had new solar panels and other vital infrastructure. There have been four more such missions. The ultraviolet spectrograph (COS) installed in 2009 is up to 20 times more sensitive than the previous ones. That is equivalent to increasing the mirror diameter from 2.4 metres (the largest that could fit inside the Space Shuttle bay) to more than 10 metres (larger than any telescope yet built).

Like Sullivan, Gorman was fascinated by space as a child, inspired by dark skies over the Australian countryside. But in common with many women at the time, she was discouraged from becoming an astrophysicist. Instead, she earned a PhD in archaeology and worked as a consultant documenting Indigenous heritage sites in her home country. But her gaze

JSC/NASA



Astronaut Michael Good works on maintaining the Hubble Space Telescope.

frequently turned skywards. Eventually, she applied her training to space exploration, regarding even the lowliest "space junk" as an important part of the historical record.

As the "Dr Space Junk" of her book's title, Gorman writes about how we should protect our space legacy. She describes the reef of orbital detritus around our planet, including satellites, alive and dead, embedded in a sea of discarded hardware and debris from space collisions (deliberate and otherwise), as well as receding planetary probes and equipment abandoned on the Moon.

She draws parallels between archaeological investigations that find ancient artefacts on Earth and missions to catalogue objects that are merely decades old circling above it. At times, her ideas can seem fanciful – as when she discusses shadows left by footprints on the lunar surface, or speculates about future civilizations finding spacecraft beyond the Solar System. But for the most part, the book made me think fresh thoughts.

Gorman reminds us how fragile our access to space is. Orbital debris alone poses a risk to every newly launched spacecraft. She warns of the need for nations to cooperate in preserving and protecting the space environment, and points out the moral responsibility of space-faring nations to deal on an equal basis with those that are not.

Both Sullivan and Gorman envision a future in which astronauts, and possibly ordinary citizens, live and work in space regularly. In that world, it will be normal to site telescopes in stable orbits at L2 (the second Lagrange point, which circles the Sun in tandem with the Earth–Moon system) and to upgrade them regularly. Hubble has seen a few monster galaxies as they were early in the evolution of the Universe. In future, more sensitive instruments, such as the James Webb Space Telescope, might see much smaller early galaxies and possibly even the first stars.

"Both Sullivan and Gorman envision a future in which astronauts live and work in space regularly."

To read these two books is to marvel at what we have achieved in our nascent efforts to inhabit space, and to recognize that we have barely begun that quest. Many popular treatments of space travel, including the films *Apollo 13* (1995) and *First Man* (2018), have framed it as competitive derring-do. Sullivan and Gorman focus more on our common interests, as humans, in knowledge and cooperation. They invite us to think anew about the legacy and the future of space.

Meg Urry is Israel Munson professor of physics and astronomy at Yale University, and director of the Yale Center for Astronomy and Astrophysics in New Haven, Connecticut. She uses the Hubble, Chandra and Spitzer space telescopes in her research on black holes.
e-mail: meg.urry@yale.edu

ALVARO YBARRA ZAVALA/GETTY

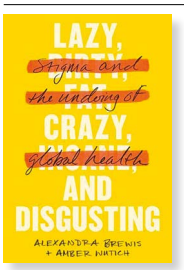
A person with HIV in the Mae Tao Clinic in Thailand.

How stigma subverts public health

A hard-hitting study exposes the devastating effects of shame and discrimination. **By Julie Pulerwitz**

As a public-health researcher working on HIV around the globe, I have seen the devastation that stigma can cause. It leads to people being shunned and isolated, and discriminated against in health care, at work and at school. And it inhibits them from accessing life-saving services and medications. As a new book by medical anthropologists Alexandra Brewis and Amber Wutich convincingly argues, stigma strips people of dignity and exacerbates the already-difficult circumstances of the poorest and most vulnerable. It can itself have major impacts on health, such as depression and even suicide.

Brewis and Wutich work in low-income countries. Their book's title, *Lazy, Crazy, and Disgusting*, highlights their areas of focus: obesity, mental illness and community sanitation. The authors focus on detailed, qualitative case studies in diverse arenas. These demonstrate three things: how stigma arises and affects the most marginalized; why stigma is so difficult to combat; and how public-health efforts can unwittingly fuel it.



Lazy, Crazy, and Disgusting: Stigma and the Undoing of Global Health
Alexandra Brewis & Amber Wutich, Johns Hopkins University Press (2019)

It is this third issue — the unintended consequences of big campaigns — that forms their main argument. And it is a rarely heard and compelling one. When, for example, the US public-health community framed smokers as putting others at risk of getting cancer from 'second-hand' smoke, the messages hit home. Social norms regarding whether it was acceptable to smoke changed, and many smokers were motivated — and managed — to quit. But there were also negative consequences. Smokers were blamed for their addiction, and people with smoking-related diseases (even those who had never smoked) were often castigated as bringing their conditions on themselves. Meanwhile, tobacco companies escaped criticism.

Similarly, the authors demonstrate how concerns about the effects of obesity — such as diabetes and cardiovascular disease — have led to fat-shaming, depression and more. Yet obesity is strongly linked to socio-economic

circumstances, and a lack of access to high-quality foods such as fresh fruit and vegetables. It is counterproductive and unfair to blame individuals. As obesity and associated conditions become increasingly prevalent across the global south, Brewis and Wutich caution against a snowball effect of harmful messaging and impacts.

The book is less strong on ways forward. I second the authors' calls for increased awareness among health practitioners, for tracking of stigma levels and for policy to be evidence-based. But, in my view, a more comprehensive and nuanced response is needed. There are important distinctions between, for example, public-health measures to reduce people's internalized feelings of blame and shame, and legislative efforts to minimize 'enacted' stigma – that is, instances of discrimination. Internalized stigma might lead to depression, and those who experience it might benefit from counselling, say. By contrast, human-rights abuses must be countered with anti-discrimination policies and laws.

Moreover, Brewis and Wutich fail to explore an important concept: intersecting stigmas. For example, a person with HIV who works in the sex industry and injects drugs might experience compounded bias and discrimination. The authors use HIV as an example of a success story in which concentrated efforts from the global health community, such as health policies and mass-media campaigns, have greatly reduced stigma.

But this is true only in some communities

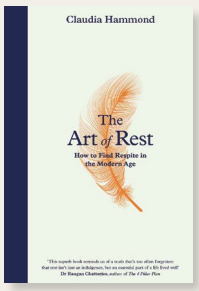
"I second the authors' calls for increased awareness among health practitioners."

– especially in high-income countries where living with HIV has been transformed into a chronic illness through the use of antiretroviral medications. These medications are often not reaching the most vulnerable, and in many contexts – for example, where drug users are criminalized and struggle to access health care – intersecting stigmas remain rampant.

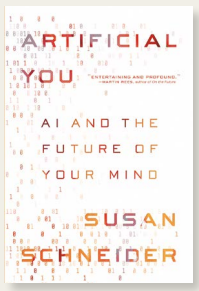
This engaging book nevertheless fills a significant gap in the literature by providing a wake-up call to scholars and practitioners unfamiliar with the topic. And it reminds me that we should all be working together to avoid any unintended consequences of promoting health.

Julie Pulerwitz directs the HIV and AIDS programme at the Population Council, an international non-profit organization based in New York City.
e-mail: jpulerwitz@popcouncil.org

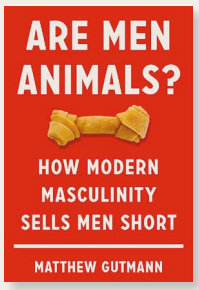
Books in brief



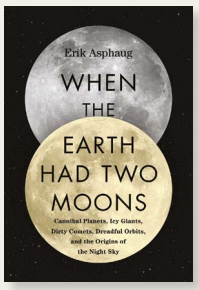
The Art of Rest
Claudia Hammond Canongate (2019)
In 2014, journalist Claudia Hammond, presenter of BBC Radio 4's *All in the Mind*, joined a group studying rest at London's Wellcome Collection. She proposed a radio survey called the Rest Test. Responses from 18,000 people in 135 countries yielded a top ten of restful activities, and they inspire the titles of her informative chapters interlacing findings from dozens of studies. Intriguingly, the top five are largely solitary. Number one is reading, which "not only allows us to escape other people, but simultaneously provides us with company", she notes.



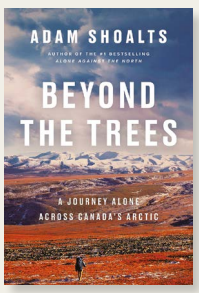
Artificial You
Susan Schneider Princeton University Press (2019)
Artificial intelligence (AI) technology will raise increasingly difficult ethical issues, argues philosopher, cognitive scientist and self-confessed technotopian Susan Schneider in this demanding dialogue between philosophy and science. How would you feel, she begins speculatively, about purchasing a "Hive Mind" — a brain chip permitting you to experience the innermost thoughts of your loved ones? That presumes, however, that future AI can capture consciousness with computation — which she argues is unlikely.



Are Men Animals?
Matthew Gutmann Basic (2019)
Anthropologist Matthew Gutmann has spent 30 years exploring concepts of masculinity across the United States, Latin America and China. "We place unreasonable trust in biological explanations of male behaviour," he argues in this wide-ranging book, which discusses US mass killings by men, Donald Trump's presidency and much more. Yet, he contends, there have been no major discoveries of a link between testosterone and aggression since 1990, despite a boom in scientific articles on the topic.



When the Earth Had Two Moons
Erik Asphaug Custom House (2019)
The days of the week are named after bodies in the Solar System and a diverse mix of Norse and Roman deities. So notes Erik Asphaug, a planetary scientist who is part of the team behind two lunar and planetary NASA missions. But if the planets were born out of material orbiting the Sun, like raindrops condensing from a cloud, why do they differ so much in structure and chemical composition? This detailed book assesses the astronomical and geological evidence on the origin of planetary diversity.



Beyond the Trees
Adam Shoalts Allen Lane (2019)
In 1967, the centennial of Canada's confederation, ten teams of canoeists paddled from central Alberta to Montreal. In mid-2017, to mark the 150th anniversary, explorer, historian and geographer Adam Shoalts travelled across the Canadian Arctic by canoe and on foot. His journey took him from the Yukon to Nunavut: across the terrestrial world's largest expanse of wilderness outside Antarctica. It proved less stressful than his normal "modern, hyper-connected world", he avers in his engaging, hazard-strewn account. **Andrew Robinson**

Comment

HARIANDI HAFID/SOPA/ZUMA WIRE

A collapsed building in the city of Palu in Sulawesi, Indonesia, after a magnitude-7.5 earthquake hit the region in September 2018.

Disaster-zone research needs a code of conduct

JC Gaillard & Lori Peek

Study the effects of earthquakes, floods and other natural hazards with sensitivity to ethical dilemmas and power imbalances.

A magnitude-7.0 earthquake rocked Anchorage, Alaska, in late November 2018. Roads buckled and chimneys tumbled from rooftops. Business operations were disrupted. Schools were damaged across the district. This was the largest earthquake to shake the region in a generation, and there was much to learn. What was the state of the infrastructure? Might further quakes occur? How did people respond? Teams of scientists and engineers from across the United States mobilized to conduct field reconnaissance in partnership with local researchers and practitioners. These efforts were coordinated through the clearing house set up by the Earthquake Engineering Research Institute in Oakland, California, which provided daily in-person and online briefings, as well as a web portal for sharing data.

But researchers are not always so welcome in disaster zones. After the deadly Indian Ocean earthquake and tsunami on 26 December 2004, hundreds of academics from countries including Japan, Russia, France and the United States rushed to the region to collect perishable data. This influx of foreign scientists angered and fatigued some locals; many declined researchers' requests for interviews. The former governor of Aceh province, Indonesia, where more than 128,000 people died, described foreign researchers as "guerrillas applying hit-and-run tactics"[1].

Yet research on tsunami propagation and people's response to the event has led to improved warnings and emergency-response plans.

When, on 28 September 2018, an earthquake and tsunami hit the Indonesian island of Sulawesi, dozens of researchers found themselves unable to enter the country[2]. Indonesian law now requires foreign scientists to obtain a special visa before they can begin research. Data-collection protocols must be submitted to the government in advance and projects must have an Indonesian partner. Violators could face criminal charges and even prison.

This incident has inflamed a smouldering debate among disaster researchers. Some scholars argue that stringent administrative protocols violate researchers' rights and prevent the collection of crucial, potentially life-saving, data[3]. Others counter that such procedures protect survivors and preserve the integrity of local scientific efforts. For instance, concerns over studies placing undue burdens on overwhelmed groups — including grieving schoolchildren — led New Zealand to impose a moratorium on social-science research after the 2011 Christchurch earthquake[4].

Here we argue that disaster research needs a culture shift. As in other branches of study involving human participants, ethical concerns should have the same primacy as research questions[5]. We call on the United Nations Office for Disaster Risk Reduction (UNDRR) to put forward a researcher-driven ethical code of conduct. This should advance disaster research, making it scientifically rigorous as well as locally and culturally grounded. After all, the UNDRR has a mandate "to ensure synergies among … regional organizations and activities in socio-economic and humanitarian fields".

Moral hazard

Researchers working in disaster zones, with people whose culture might be different from their own, need to know how to interact with survivors as well as local officials and scholars, without adding to those people's problems.

There is no universal definition of ethical behaviour, and only a handful of countries have ethically informed guidelines for post-disaster research. In New Zealand, guiding principles from the Natural Hazards Research Platform advise that researchers must "avoid creating unnecessary anxiety by speculating to locals". The Philippines allows research on the trauma caused by disasters only in exceptional cases, such as when affected people want to share their feelings as a way to process the event. Brazil, like Indonesia, requires all researchers working in the country to have a special visa and an established local connection.

University ethics committees and national ethical review boards are unable to fill the gap. They tend to focus on studies in medicine and social sciences that involve human participants. They have little to say on how to investigate a collapsed building or a compromised coastal landscape. Yet studies by engineers or natural scientists have participants, too: local residents, scholars, guides and interpreters. Tsunami researchers might need to ask coastal dwellers about the height of waves; structural engineers assessing a collapsed stairwell might question the building's occupants about how they escaped.

> **"Ethical concerns should have the same primacy as research questions."**

Towards a code of conduct

Researchers equipped with an 'ethical toolkit' are better able to help affected populations[6] without causing harm. Following the earthquake that struck Luzon island in the Philippines in April this year, research was coordinated by academics based in nearby Manila. They provided support deemed appropriate by those affected. A code of conduct could build on such successes and should consider the following three principles.

Have a clear purpose. Researchers should collectively identify knowledge gaps that future studies will fill. They should partner with affected people to establish emergent research priorities in dealing with a disaster. Such collaborative engagement can help to clarify where and when researchers will go into the field, what they will study, and who should be on the team. For example, psychologists and anthropologists might study and support local coping mechanisms; historians and civil engineers might collaborate to examine and promote resilient traditional architectural features when rebuilding homes in cyclone-affected areas.

The needs of local people should be central[7]. Too often, research is driven by media coverage and politics. Disasters in heavily populated areas receive the most attention, but the cumulative impacts of smaller events can be just as devastating. For example, after the massive Nepal earthquake in April 2015, the impacts on infrastructure and the quality of shelters were widely studied, and aid donors gave millions of dollars to rebuild parts of Kathmandu. Yet in rural western Nepal, hundreds of villages cope with floods and landslides each year, unnoticed by the outside world.

A researcher code can help to redress the balance. For example, the Philippines requires that post-disaster projects demonstrate how they will meet the priorities of affected communities. New Zealand encourages researchers to defer collecting data unless the information will support responders. More relevant research could provide the evidence to inform and direct recovery funding to where needs really lie.

Respect local voices. Wealthy countries account for most disaster scholarship and funding. For example, more than 90% of articles published following Hurricane Katrina, which hit the southern United States in 2005, were by US researchers[8]. By contrast, fewer than 5% of publications on the 2010 Haiti earthquake were led by authors based in the country (see 'Unequal partners').

Similarly, 84% of articles published between 1977 and 2017 in *Disasters*, the flagship journal in the field, were led by authors based in countries of the Organisation for Economic Co-operation and Development (OECD). Yet 93% of the people killed by large disasters over the same period lived in non-OECD countries, according to the EM-DAT disaster database[9].

Outside researchers — who have not had their lives disrupted by disaster — are positioned to seek funding and might overlook local work and partners. After Hurricane Katrina in 2005, local experts in urban poverty, affordable housing and coastal land loss were passed over for grants[10]. And local and external priorities might differ. In 2011, following the Joplin tornado in Missouri, outside academics assessed damage to infrastructure. By contrast, locally based researchers were eager to learn how to support emotional health after witnessing a rise in post-traumatic stress in children and adults[11]. Both are important topics, but funding streams do not always follow local desires.

An understanding of local languages, policies and practices is essential and can improve response and speed recovery. After Katrina, 'culture brokers' helped survivors to make sense of government documents so that they could access aid quickly[12]. Nonetheless, much disaster research is still framed by narrow world views. Concepts such as vulnerability and resilience do not necessarily translate well[13]. Even where equivalent terms exist, they might be felt to be irrelevant, because natural phenomena such as cyclones and floods are not always seen as hazards. In some religious

traditions, volcanic eruptions are thought to reflect the emotions of deities, for instance. A lack of recognition of this nuance can affect the outcomes of risk-perception research as well as early-warning processes.

More discussions between disaster researchers inside and outside affected areas would shed light on these issues and could inform a more holistic research agenda. The Geotechnical Extreme Events Reconnaissance Association's ethics protocol might serve as a starting point. It encourages engineers to adhere to "high standards of professionalism" and to be "respectful of local customs, traditions, privacy, and rights of affected individuals" (see go.nature.com/32kptno). Government agencies, companies and non-governmental organizations should also be involved in such conversations, given that they are increasingly engaged in post-disaster data collection[14].

Coordinate locals and outsiders. Projects that are uncoordinated can become irrelevant or redundant, and might overwhelm local people and responders. In 2013, survivors of Typhoon Yolanda (also known as Haiyan) in Tacloban in the Philippines were deluged with questionnaires, when their immediate concerns were to secure housing, food, clothing and education.

After Hurricane Harvey in the United States in 2017, officials at emergency operations centres struggled to decipher the credentials of dozens of researchers who descended on Houston, Texas, requesting access. Emergency managers also had to spend precious time revising researchers' survey questions to put them in a local context.

Foreign scientists sometimes approach local researchers to serve as translators or assistants. These locals have little power to direct the research strategy, even though their insights are valuable. They might feel unable to be critical even when they know the questions are wrong-headed. Even when they make substantial contributions, they might still be relegated to co-authorship — or no authorship — rather than being listed as the primary author.

Incoherent data and findings might confuse authorities and delay decisions. Volcanologists still argue about exactly when local communities should be evacuated. To help, the International Association of Volcanology and Chemistry of the Earth's Interior has produced guidelines on the roles and responsibilities of local and outside scientists, local authorities and the media.

> **"Much disaster research is still framed by narrow world views."**

Local researchers need to be identified quickly in a crisis. As a start, the Social Science Extreme Events Research (SSEER) network has produced a global map of social scientists who study hazards and disasters (see go.nature.com/2qfwezc). Regional SSEER councils ensure that those researchers remain involved after the event.

First steps

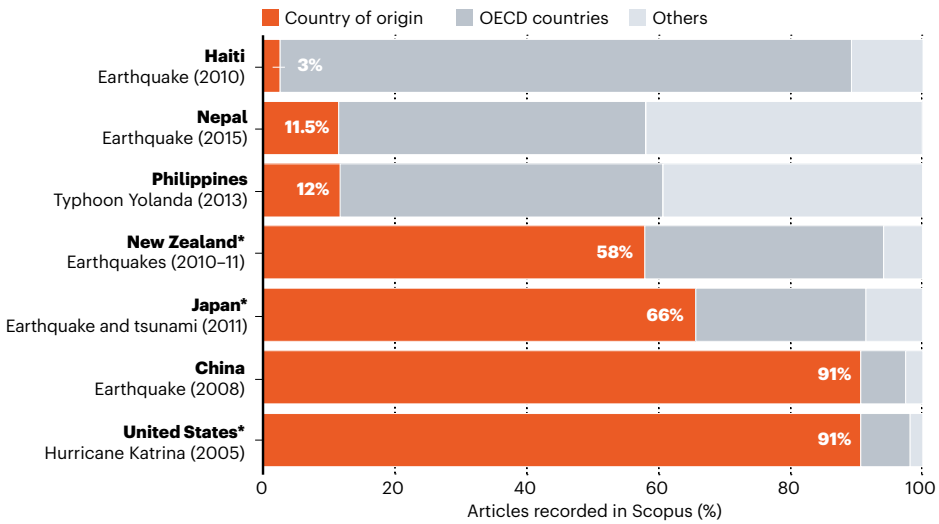
Discussions regarding a shared code of conduct could start through collaborative disaster-research initiatives that are under way worldwide. These have established strong coordinating structures and forums for information sharing, and include those in Latin America, Africa, the European Union and the Asia-Pacific region. They could also build on disaster-response initiatives from the medical sciences[15,16].

The US National Science Foundation (NSF) now supports several extreme-events reconnaissance and research networks. These advance coordination and set scientific agendas in geotechnical and structural engineering, social sciences, near-shore systems, operations and systems engineering, and interdisciplinary science. The NSF-funded CONVERGE initiative (of which L.P. is the principal investigator) brings together leaders from these networks and major NSF facilities to support the development of guidance and data-sharing by hazards and disaster researchers. Other resources, including a set of free online training modules, are also available. These NSF initiatives are open to researchers globally, but they are led by researchers at US institutions.

Most countries do not provide ethical guidance for researchers, and universities have widely varying standards for the protection of study participants. The UNDRR is a trusted convener of scientists and practitioners globally. It could serve as a focal point for the development and implementation of an ethical code of conduct for researchers in disaster zones. As disasters unfold around the globe, the need for such a code of conduct becomes ever more urgent.

UNEQUAL PARTNERS

Authorship of papers on disaster research can be dominated by researchers outside the country affected, meaning that local expertise might be overlooked.



* 'OECD countries' category excludes the affected member country of the Organisation for Economic Co-operation and Development.

SOURCE: SCOPUS/ADAPTED FROM REF. 8

The authors

JC Gaillard is an associate professor in the School of Environment at the University of Auckland, New Zealand; and extraordinary professor in the Unit for Environmental Sciences and Management, North-West University, Potchefstroom, South Africa. **Lori Peek** is professor of sociology and director of the Natural Hazards Center, University of Colorado Boulder, USA; and principal investigator of the US National Science Foundation-supported CONVERGE, SSEER and ISEEER Initiatives.
e-mails: jc.gaillard@auckland.ac.nz; lori.peek@colorado.edu

1. Missbach, A. *Crit. Asian Stud.* **43**, 373–398 (2011).
2. *Nature* **562**, 317–318 (2018).
3. Kendra, J. & Gregory, S. in *Disaster Research and the Second Environmental Crisis* (eds Kendra, J., Knowles, S. G. & Wachtendorf, T.) 319–341 (Springer, 2019).
4. Beaven, S., Wilson, T., Johnston, L., Johnston, D. & Smith, R. *Earthq. Spectra* **32**, 713–735 (2016).
5. Guillemin, M. & Gillam, L. *Qual. Inq.* **10**, 261–280 (2004).
6. Browne, K. E. & Peek, L. *Int. J. Mass Emerg. Disasters* **32**, 82–120 (2013).
7. Chambers, R. *Rural Development: Putting the Last First* (Longman, 1983).
8. Gaillard, J. C. *Disasters* **43**, S7–S17 (2019).
9. Gaillard, J. C. & Gomez, C. *Jàmbá: J. Disaster Risk Stud.* https://doi.org/10.4102/jamba.v7i1.120 (2015).
10. Barber, K. & Haney, T. J. *Sociol. Spectrum* **36**, 57–74 (2016).
11. Houston, J. B. *et al. PLoS Curr.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4639320 (2015).
12. Browne, K. *Standing in the Need* (Univ. Texas Press, 2015).
13. Bankoff, G. *Disasters* **25**, 19–35 (2001).
14. Miller, A. *et al. Int. J. Environ. Res. Public Health* **13**, 676 (2016).
15. Sumathipala, A. *et al. Asian Bioeth. Rev.* **2**, 124–142 (2007).
16. Collogan, L. K., Tuma, F., Dolan-Sewell, R., Borja, S. & Fleischman, A. R. *J. Trauma. Stress* **17**, 363–372 (2004).

Correspondence

Chile: democratizing policies online

Chile became a political hotbed in a matter of days last month. Amid the chaos, people demanded reforms to the country's privatized pension and health-care systems, a new constitution, and punitive measures for tax dodgers and companies involved in price-fixing. But their voices need to be aggregated if they are not to be lost in the din of rallies or fragmented into thousands of tweets.

To this end, we created the experimental platform Chilecracia, using crowdsourcing methods already validated in academia. Examples include MIT's Place Pulse (P. Salesses *et al*. *PLoS ONE* **8**, e68400; 2013) and Moral Machine (E. Awad *et al*. *Nature* **563**, 59–64; 2018). Chilecracia pairs policy proposals and asks: "What would you prioritize?" Within 10 days, more than 120,000 people had indicated at least one preference, amounting to more than 7 million votes. The data are helping us to compile detailed networks of policy preferences (see chilecracia.org).

Chilecracia is being updated weekly with the help of a team of policy experts. We have received requests from several countries to deploy regional, organizational and national instances of the system. Our findings add to the growing literature on such surveys (see go.nature.com/2qiwoja) and offer insight into online crowdsourced participation systems in politically active situations.

César A. Hidalgo University of Toulouse, France.
cesifoti@gmail.com

C.A.H. declares competing financial interests; see go.nature.com/2nxscdc.

MARTIN BERNETTI/AFP/GETTY

Huge demonstrations have swept through Chile since mid-October.

Chile: science could tackle social unrest

Finding solutions to Chile's current social turmoil will demand efforts from every sector, including the research community (see *Nature* **575**, 265–266; 2019). Our contribution will depend on government support for a more ambitious, participatory policy for science, technology and innovation.

In the past, Chile's science policy has focused on boosting productivity and economic growth. However, the problems highlighted by the latest social unrest are unlikely to be solved just by increasing gross domestic product (see also P. A. Besnier *Nature* **511**, 385; 2014).

Creating a more comprehensive science and innovation policy will also require a new, improved way of doing politics. More players must be involved, including citizen representatives, to allow their perspective to optimize future policies for society's well-being.

Pablo Astudillo Besnier Autonomous University of Chile, Santiago, Chile.
pablo.astudillo@uautonoma.cl

Don't bury hidden treasure

There is much to applaud in the EON-ROSE (Earth-system Observing Network-Réseau d'Observation du Système Terrestre) project to understand Canada's geology and to find geothermal energy (see *Nature* **574**, 463–464; 2019). But as your heading 'Hidden treasure' indicates, there are also commercial implications in prospecting for mineral deposits.

The boundaries between science, economics and politics are often invisible. Scientific ventures should always be open about their potentially commercial objectives, as in the EON-ROSE case, particularly if they are financed by funding agencies and indirectly by taxpayers.

In my view, scientists should reject outright any subsidies that support the search for and development of carbon-based energy and of environmentally destructive mineral extraction.

Talan İşcan Dalhousie University, Halifax, Canada.
talan.iscan@dal.ca

Writer's secret? No interruptions

Your piece 'Day in the life of a 24-hour global news factory' revived memories of my only visit to the *Nature* office (see go.nature.com/2k9scd1). The occasion was in the early 1980s, when John Maddox was in his second term as editor-in-chief and he invited me to report news stories from India for the journal. He wrote many himself, one of which included an interview with Indira Gandhi, the country's prime minister at the time (see *Nature* **308**, 582; 1984).

When I met the great man again, he was in hospital being treated for a leg injury. He passed his days there avidly reading, writing and editing as usual. I enquired after his secret of being able to write such insightful editorials on a host of topics – ranging from physics to philosophy – week after week. His reply was that he always firmly shut his office door, allowing no phone calls or other interruptions until he had completed his next editorial.

Killugudi Jayaraman Bangalore, India.
killugudi@hotmail.com

HOW TO SUBMIT

Correspondence may be submitted to correspondence@nature.com after consulting the author guidelines and section policies at go.nature.com/cmchno.

News & views

Palaeontology

Fossil ape hints at how bipedal walking evolved

Tracy L. Kivell

Approximately 11.6-million-year-old fossils reveal an ape with arms suited to hanging in trees but human-like legs, suggesting a form of locomotion that might push back the timeline for when walking on two feet evolved. **See p.489**

Ever since Charles Darwin's work provided the basis for understanding human evolution, there have been long-standing questions regarding when, why and how our early human ancestors begin to walk on two feet. The commitment to terrestrial bipedalism, characterized by skeletal adaptations for walking regularly on two feet, is a defining feature that enables the assignment of fossils to the hominin lineage – which comprises all species more closely related to humans than to chimpanzees (*Pan troglodytes*) or bonobos (*Pan paniscus*), our two closest living relatives. On the basis of fossil findings, some of which are more controversial than others[1,2], the answer to the 'when' question is thought to be between 7 million and 5 million years ago at the end of the Miocene epoch (which lasted from about 23 million to 5 million years ago).

Answering the questions of why and how hominin bipedalism evolved depends a lot on what kind of locomotion was being used before terrestrial bipedalism evolved. Did it evolve from an ancestor that lived mainly in trees, or were these ancestors already walking on all fours on the ground and subsequently evolved to stand up and walk on two feet? On page 489, Böhme *et al.*[3] report the discovery of an ape species called *Danuvius guggenmosi* from the middle of the Miocene. This species moved around in a previously unknown way, which the authors suggest could provide a model for the type of locomotion from which hominin bipedalism evolved.

Questions about the origin of hominin bipedalism and how the last common ancestor of humans, chimpanzees and bonobos might have moved are conventionally addressed using either a top-down or a bottom-up approach (Fig. 1). Darwin[4] and many palaeoanthropologists favoured the top-down approach, examining living primates, particularly the great apes, for clues to how bipedalism evolved[5,6]. African apes – chimpanzees, bonobos and gorillas (of the genus *Gorilla*) – go into the trees to eat, sleep and when they need protection, but spend most of their time on the ground, using their knuckles for walking. Given our close genetic relationship to these apes, and because we also share certain features of our hands and feet with them, some have argued that hominin bipedalism evolved from a knuckle-walking ancestor[5], or a more generalized quadruped lacking knuckle-walking specializations[7], that divided its time between the ground and the trees. By contrast, others have noted that the way that orangutans (of the genus *Pongo*) move bipedally in trees, and the mechanical similarities between how apes use their legs for climbing and how humans use theirs for walking, suggest that bipedalism evolved from an ape ancestor that was previously committed to life in the trees[6,8].

Although logical, this top-down approach is constrained, as Darwin acknowledged[4], to examining evidence from the few remaining living ape species. However, one of the earliest potential hominins for which we have the most fossil evidence – the approximately 4.4-million-year-old *Ardipithecus ramidus* – is argued to be distinctly unlike living great apes in its anatomy, which suggests that the African apes and Asian orangutans we know today are actually quite specialized in their locomotor behaviours compared with their earlier ancestors[7]. Each living ape species is a result of its own long, evolutionary history,

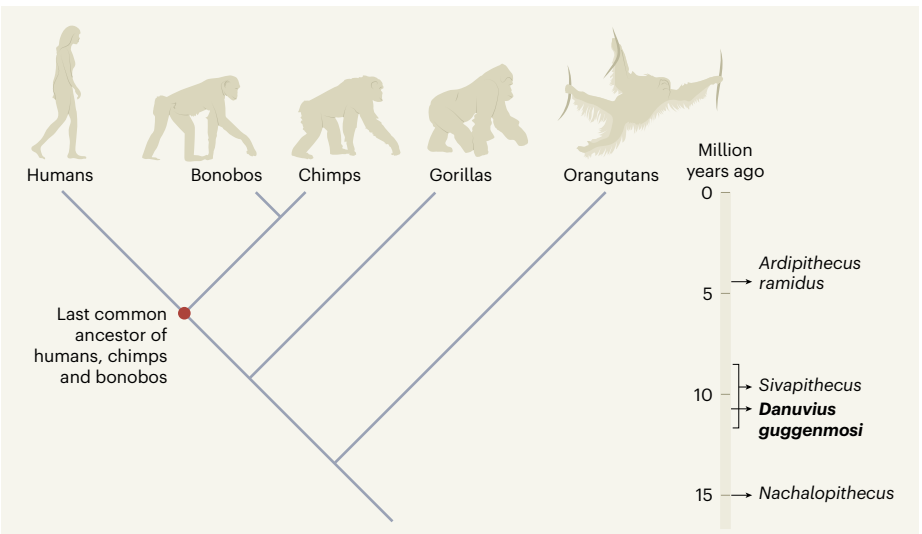


Figure 1 | The evolution of bipedalism. In the branch of the evolutionary tree that splits from our last common ancestor with chimpanzees (*Pan troglodytes*) and bonobos (*Pan paniscus*), humans and our extinct hominin relatives have a skeleton adapted for regular walking on the ground using two feet. A top-down approach to assessing how our early ancestors might have evolved bipedalism focuses on possible modes of ancestral locomotion by considering how living great apes move around. For example, African apes – chimps, bonobos and gorillas (of the genus *Gorilla*) – use knuckle-walking more frequently on the ground than in trees, and these apes and orangutans (of the genus *Pongo*) also climb and use suspensory locomotion in trees. However, fossils of the ancient potential hominin *Ardipithecus ramidus* suggest that living apes might have evolved quite specialized locomotion compared with their earlier ancestors. A bottom-up approach focuses instead on ancient ape fossils that pre-date our last common ancestor, such as those of the genus *Nacholapithecus* or *Sivapithecus*. However, the clues uncovered from such fossils can be difficult to interpret. Böhme *et al.*[3] present fossils of a previously unknown ape called *Danuvius guggenmosi*, which the authors suggest provides a good model for the type of locomotion from which bipedalism might have evolved. The branch-point timings shown are approximate.

and, in the case of African apes, one that we often forget because there is so little fossil evidence of it. This absence of fossil information to reveal how African apes evolved makes questions about the nature of our common ancestor even trickier to answer.

Other palaeoanthropologists address the question of bipedal origins from a bottom-up approach instead, looking to the approximately 30 genera of fossil apes that have been identified from the Miocene of Africa, Asia and Europe as potential models for what our last common ancestor might have looked like[2,7,9]. However, these apes show a hotchpotch of skeletal adaptations, with features found in combinations that are unlike anything we see in living primates, and that often leave us guessing about how these animals moved around and how much time they spent in trees or on the ground.

For example, a genus of fossil ape called *Nacholapithecus* had a monkey-like body but unusually large forelimbs and long toes, whereas another ape genus, *Sivapithecus*, had an orangutan-like face, an ape-like shoulder, and a monkey-like elbow and pelvis[10,11]. Such characteristics suggest odd combinations of arboreal suspension (hanging from tree branches), quadrupedal movements and body postures that are difficult to imagine today, and which make it hard to interpret these creatures' probable locomotion patterns[10].

Böhme and colleagues add to this amazing Miocene diversity by presenting approximately 11.6-million-year-old fossils of *D. guggenmosi*. The authors interpret the shape of the *D. guggenmosi* fossils as indicating a type of previously unknown movement that they term extended limb clambering, which combines adaptations of both suspension in the trees and bipedal locomotion. This makes it a good possible model of locomotion for the last common ancestor.

The teeth of *D. guggenmosi* identify it as belonging to a group of fossil ape species called dryopithecins that have been found from the mid- to late Miocene in Europe and that some consider to be ancestral to African apes[9]. Living African ape species inhabit the equatorial region of Africa, but, during certain times of the Miocene, many ancestral great apes were living throughout Europe and Asia and migrating both to and out of Africa. Some researchers suggest that the dryopithecins show features found in chimps and gorillas today and therefore make good candidates for the ancestors of living African apes[9]. The *D. guggenmosi* skeleton is unique compared with other dryopithecin specimens, both in its preservation of two, almost complete, limb bones — an ulna (a forearm bone) and a tibia (a leg bone) — and in the combination of characteristics it displays. Böhme *et al.* focus their attention on a baboon-sized and probably male partial skeleton. As well as the ulna and tibia, the skeleton includes some vertebrae, a partial thigh bone (femur), and hand and foot bones.

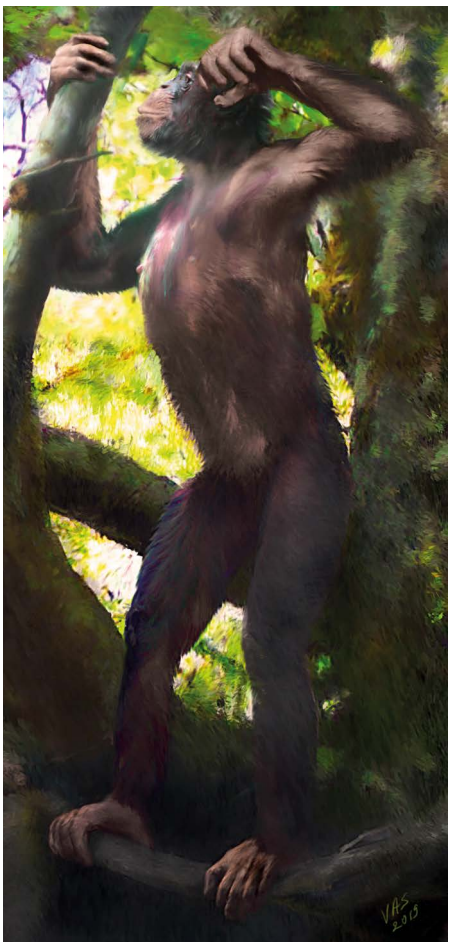


Figure 2 | *Danuvius guggenmosi*. Böhme and colleagues instructed the artist Velizar Simeonovski to make an illustration of what this species might have looked like.

VELIZAR SIMEONOVSKI

The length of the ulna relative to the tibia shows that the forearm of *D. guggenmosi* was long relative to the leg, similar to a bonobo's form. Combined with a flexible elbow and hand bones indicating a powerful, grasping thumb and curved fingers, the forelimb has the telltale signs of arboreal suspension found in all living great apes.

"The newly discovered ape species might have walked flat-footed on branches."

However, the lower limb of *D. guggenmosi* tells a different story, and one that is more reminiscent of human lower limbs than of those of other great apes. The shape of the joints of the femur and tibia suggests the use of extended (upright) hip and knee postures that differ from the bent hips and knees that living African apes use when they occasionally walk bipedally on the ground or in trees. The top of the tibia is reinforced, and the ankle joint is stable, properties that are adaptations for resisting the higher load placed on the lower leg when moving on two limbs instead of four. But the foot has a long, robust big toe that would be good for grasping, suggesting that *D. guggenmosi* might have walked flat-footed on branches (Fig. 2). Whether or not it regularly walked bipedally on the ground is less clear.

Together, the mosaic features of *D. guggenmosi* arguably provide the best model yet of what a common ancestor of humans and African apes might have looked like. It offers something for everyone: the forelimbs suited to life in the trees that all living apes, including humans, still have; lower limbs suited to extended postures like those used by orangutans during bipedalism in the trees[8]; and further specialization of such features of the lower limbs in humans to enable habitual terrestrial bipedalism.

If it is accepted that the locomotor behaviours observed in living great apes and humans evolved from an ancestor that used extended limb clambering, this would answer the question of what kind of early locomotion underlies our bipedal origins. And that would get us closer to answering why and how our human ancestors became less dependent on life in the trees and fully embraced two-footed terrestrial locomotion. Until more fossil evidence of how African apes evolved is found, a bottom-up approach from the Miocene is probably our best means of deciphering the evolution of one of our most defining human features.

Tracy L. Kivell is at the School of Anthropology and Conservation, University of Kent, Canterbury CT2 7NR, UK, and at the Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.
e-mail: t.l.kivell@kent.ac.uk

1. Wood. B. & Harrison, T. *Nature* **470**, 347–352 (2011).
2. Almécija, S. *et al. Nature Commun.* **4**, 2888 (2013).
3. Böhme, M. *et al. Nature* **575**, 489–493 (2019).
4. Darwin, C. *The Descent of Man* (Murray, 1871).
5. Richmond, B. G., Begun, D. R. & Strait, D. S. *Am. J. Phys. Anthropol.* **116**, 70–105 (2001).
6. Crompton, R. H., Seller, W. I. & Thorpe, S. K. S. *Phil. Trans. R. Soc. B* **265**, 3301–3314 (2010).
7. Lovejoy, C. O. *Science* **326**, 74 (2009).
8. Thorpe, S. K. S., Holder, R. L. & Crompton, R. H. *Science* **316**, 1328–1331 (2007).
9. Begun, D. R., Nargolwalla, M. C. & Kordos, L. *Evol. Anthropol.* **21**, 10–23 (2012).
10. Ward, C. V. in *Handbook of Paleoanthropology* (eds Henke, W. & Tattersall, I.) 1363–1386 (Springer, 2015).
11. Morgan, M. E. *et al. Proc. Natl Acad. Sci. USA* **112**, 82–87 (2015).

This article was published online on 6 November 2019.

Structural biology

Snapshots of a genetic cut-and-paste

Orsolya Barabas

Transposase proteins mediate the movement of 'parasitic' DNA segments in genomes. A series of structures of a transposase catches it in action, and highlights how these proteins evolved for use in immune systems. **See p.540**

Our genetic material is littered with parasitic DNA sequences, known as transposons, which promote their own propagation and transmission, rather than their host's[1]. Their movements (transposition) within and between genomes have profound consequences for the genetic code – sometimes leading to diseases, but also driving genetic diversity and evolution[2]. Transposase proteins mediate this movement by executing all the required chemical reactions.

Strikingly, these proteins have been repeatedly repurposed throughout evolution to produce new biological functions that benefit the host. A prime example is the vertebrate immune system, in which transposase-like RAG proteins help to assemble new genes from three pools of interchangeable DNA parts (known as V, D and J gene segments). This process is called V(D)J recombination, and equips immune cells with a diverse set of sensors that can recognize many threats[3]. On page 540, Liu *et al.*[4] report a series of structures of a transposase that is an ancestor[5] of RAG, casting light on the evolutionary history of these proteins.

Transposases must recognize several DNA sites, and then cut and join them in the proper order[6]. To understand this multi-step process, we need to visualize the structures of the molecular machinery involved at all stages, which is a major technical challenge. Liu *et al.* have now drawn on a powerful combination of two techniques – X-ray crystallography and single-particle cryo-electron microscopy – to picture several steps of transposition in remarkable detail, thereby providing a molecular 'movie' of the process.

The authors' achievements build on many years of structural studies of transposases[7–11] and RAG[12–15], providing an increasingly complete view of their functions and helping to connect the dots between the 'selfish' DNA rearrangements of transposases and the essential functions that evolved from them. Transposases are now known to have a catalytic core unit and diverse extra parts that bind to DNA or control transposase function[6]. They usually act in pairs, with each dimer holding two segments of the transposon DNA .

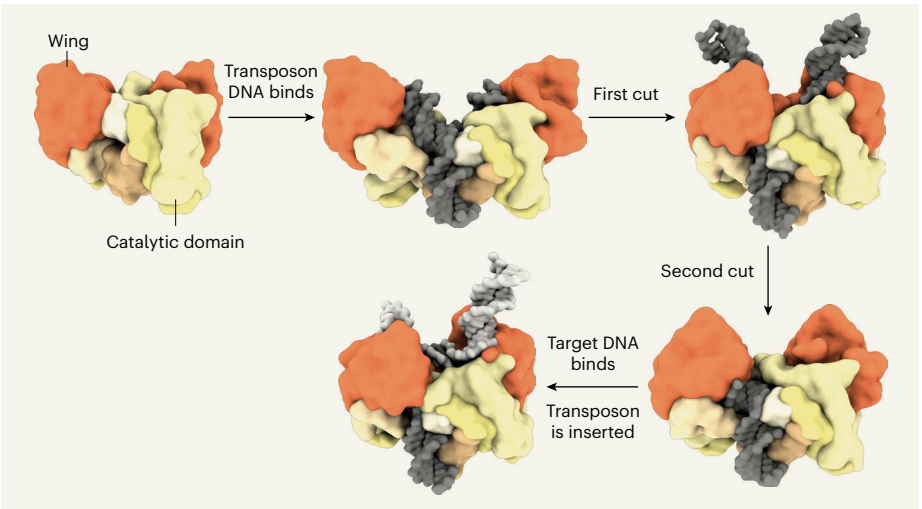
The transposase studied by Liu *et al.* mediates the movement of a transposon called Transib, and was isolated from a moth (*Helicoverpa zea*). Appropriately, the shape of the transposase complex resembles that of a moth: each 'wing' comes from one of the two transposase molecules in the complex, and is formed mainly from a protein region called the zinc-binding (ZnB) domain. The wings provide many of the interactions with the DNA, which forms the 'antennae' (Fig. 1).

As with many transposons, Transib moves by a cut-and-paste process: its transposase cleaves it out of the genome, using a catalytic core present in many transposases, and inserts it elsewhere in the genome[16]. Liu and co-workers' structures of five steps in Transib transposition now reveal the remarkable conformational changes in the protein during this process.

Perhaps most notably, the authors find that the ZnB wings move constantly: they unfurl when transposon DNA first arrives, and then close and open again during the rest of the process. This 'flapping' accompanies some impressive DNA acrobatics, which brings different DNA parts into the protein's core for ordered cleavage and joining. Remarkably, the ZnB domains help to capture not only the transposon, but also the target molecules into which Transib will be inserted – first opening to make space for the molecules, and then closing to plug the region of the target DNA into which Transib will be integrated into the core. The structures also show that the end section of the protein (the carboxy-terminal tail; CTT) contains three short α-helices that form an accordion-like structure, which connects the moving wings to the complex's 'body'. Previously reported transposase structures[6–11] have revealed similar overall features, but the movements in the Transib transposase are much more extensive.

RAG protein complexes consist of two transposase-like RAG1 proteins and two RAG2 proteins. In these complexes, ZnB is present in RAG1, but is more fixed than in the Transib transposase; and no part is present[13] that is analogous to CTT. RAG2, which is essential in V(D)J recombination but absent in most transposases, sits above the wings of RAG1 and holds a large part of the DNA[12] – much as ZnB and CTT do in the Transib transposase.

Unlike transposases, the RAG complex cuts



VLADIMIR ARINKIN & ORSOLYA BARABAS

Figure 1 | Flapping of a transposase complex. Liu *et al.*[4] report a series of structures of the transposase enzyme that mediates the movement of the Transib transposon (a parasitic genetic element) within genomes; the first structure shown was obtained using X-ray crystallography, and the others were obtained using single-particle cryo-electron microscopy. The transposase forms a dimeric complex that is roughly moth-shaped. The 'wings' unfurl to capture transposon-containing DNA, and then close again as the catalytic domain makes the first cut to cleave the transposase out of the DNA. The wings open again after the second cut, allowing target DNA (the DNA into which the transposon will be inserted) to be captured and the transposon to be inserted. The DNA sequences differ in length in some of the panels.

and removes the distinct DNA sequences found between the V, D and J gene segments, tightly coordinating the process to ensure that different types of segment are subsequently connected. In cells, RAG also largely stops the removed DNA from being reinserted elsewhere in the genome, to prevent potentially harmful changes to the genetic code. But how did these functions evolve? A logical proposal implicates RAG2, but a recent study[11] of ProtoRAG — a relative of RAG found in invertebrates that contains RAG2 but still acts as a transposase — shows that things are more complicated. Elements in both RAG1 and RAG2 help to coordinate DNA cleavage and prevent insertion.

Liu and colleagues' findings cast fresh light on the role of RAG2, showing that it carries out many of the functions of ZnB, but increases the rigidity of the whole RAG complex, compared with that of the transposase complex. It binds the DNA at the V, D and J segments more tightly than ZnB binds at the transposon, and does not undergo such large conformational changes (which can require a lot of energy, and thus reduce efficiency). This increased rigidity and tight binding might help to ensure the strict molecular coordination required for V(D)J recombination. It might also prevent release of cleaved DNA segments and/or stop the wings from reopening to accept any other DNA molecules — thereby preventing removed DNA from being reinserted elsewhere in the genome. If the wings do not open, then any incoming DNA would have to bend itself to an angle of about 150° before entering the protein, which is not easily done.

Note that Liu *et al.* were not able to directly observe the structure of the transposase in complex with intact target DNA. It therefore remains to be seen whether target DNA first binds to the transposase in a relaxed form and is then forced into a severe 150° bend. The authors also did not observe a complex in which the transposase binds intact transposon DNA such that the catalytic core is close enough to the ends to cleave them; instead, the authors observed intact transposon DNA bound with its ends away from the catalytic centre. In RAG, a large twist in the DNA is needed to position its breakpoints accurately for the cuts[14]. A similar twist might occur in Transib, but other explanations are also possible.

Efforts are now needed to define the exact functions of RAG2. Curiously, the cell-free RAG complex can readily insert excised DNA into another DNA molecule[17,18] (a target DNA). Structures of RAG with a bound target DNA must therefore be obtained — ideally, both with the intact target and after insertion. These structures will show whether the target DNA becomes as sharply bent as it does in the Transib transposase, and reveal how RAG2 affects the binding of target DNA and its insertion of excised DNA.

Other proteins might be needed to promote the function of RAG. This possibility has previously been investigated, but the availability of new structures and methods provides further opportunities for research. For example, large molecular assemblies can now be studied inside cells using a technique called electron tomography[19], and molecular interactions can be probed with advanced mass-spectrometry methods[20]. Analysis of genomic data from different species will also be helpful in identifying ancestors of RAG proteins other than ProtoRAG and the Transib transposase, and thereby exploring their evolutionary history. Such research will help to explain how parasitic genetic elements can be repurposed for crucial biological functions.

Orsolya Barabas is in the Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany.
e-mail: barabas@embl.de

1. Lander, E. S. *et al. Nature* **409**, 860–921 (2001).
2. Bourque, G. *et al. Genome Biol.* **19**, 199 (2018).
3. Jones, J. M. & Gellert, M. *Immunol Rev.* **200**, 233–248 (2004).
4. Liu, C., Yang, Y. & Schatz, D. G. *Nature* **575**, 540–544 (2019)
5. Kapitonov, V. V. & Jurka, J. *PLoS Biol.* **3**, e181 (2005).
6. Hickman, A. B. & Dyda, F. *Microbiol. Spectr.* **3**, MDNA3-0034-2014 (2015).
7. Davies, D. R., Goryshin, I. Y., Reznikoff, W. S. & Rayment, I. *Science* **289**, 77–85 (2000).
8. Richardson, J. M., Colloms, S. D., Finnegan, D. J. & Walkinshaw, M. D. *Cell* **138**, 1096–1108 (2009).
9. Hickman, A. B. *et al. Nucleic Acids Res.* **46**, 10286–10301 (2018).
10. Montaño, S. P., Pigli, Y. Z. & Rice, P. A. *Nature* **491**, 413–417 (2012).
11. Zhang, Y. *et al. Nature* **569**, 79–84 (2019).
12. Ru, H. *et al. Cell* **163**, 1138–1152 (2015).
13. Kim, M.-S. *et al. Mol. Cell* **70**, 358–370 (2018).
14. Ru, H. *et al. Nature Struct. Mol. Biol.* **25**, 732–742 (2018).
15. Kim, M.-S., Lapkouski, M., Yang, W. & Gellert, M. *Nature* **518**, 507–511 (2015).
16. Chen, S. & Li, X. *Gene* **408**, 51–63 (2008).
17. Hiom, K., Melek, M. & Gellert, M. *Cell* **94**, 463–470 (1998).
18. Agrawal, A., Eastman, Q. M. & Schatz, D. G. *Nature* **394**, 744–751 (1998).
19. Mahamid, J. *et al. Science* **351**, 969–972 (2016).
20. Smits, A. H. & Vermeulen, M. *Trends Biotechnol.* **34**, 825–834 (2016).

This article was published online on 13 November 2019.

Astrophysics

Extreme emission seen from γ-ray bursts

Bing Zhang

Cosmic explosions called γ-ray bursts are the most energetic bursting events in the Universe. Observations of extremely high-energy emission from two γ-ray bursts provide a new way to study these gigantic explosions. **See p.455, p.459 & p.464**

Astrophysical explosions known as γ-ray bursts (GRBs) can release in one second the amount of energy that the Sun will produce in its entire lifetime[1]. The emission from GRBs covers a broad stretch of the electromagnetic spectrum and occurs in two stages: the prompt-emission phase and the afterglow phase. The main emission mechanism is thought to be synchrotron radiation, whereby the gyration of energetic electrons around magnetic-field lines releases photons. Until now, emission from GRBs has been observed only at energies below 100 gigaelectronvolts (GeV). Three papers in this issue[2–4] report observations of γ-rays that have energies above 100 GeV from two bright GRBs, dubbed GRB 190114C and GRB 180720B.

The Major Atmospheric Gamma Imaging Cherenkov (MAGIC) Collaboration[2] (page 455) detected photons in the teraelectronvolt range (1 TeV is 10^3 GeV) from GRB 190114C, using the MAGIC telescopes at La Palma, Spain. The first detections started about one minute after the burst triggered NASA's two spaceborne GRB detectors: the Burst Alert Telescope on board the Swift satellite and the Gamma-ray Burst Monitor on board the Fermi satellite. The high-energy photons continued to rain down on the MAGIC telescopes for about 20 minutes, with the flux decreasing rapidly over this time. The MAGIC Collaboration and colleagues[3] (page 459) detected this GRB using several other ground-based and space-borne telescopes. When combined with the MAGIC data, this rich data set allowed the authors to model the event comprehensively and study how the TeV emission was produced.

Abdalla *et al.*[4] (page 464) detected photons of energies above 100 GeV (but below 1 TeV) from GRB 180720B, using the High Energy Stereoscopic System (HESS) array of telescopes in Namibia. Although these photons were lower in energy and fewer in number than those observed from GRB 190114C, they were detected from deep in the afterglow phase (10 hours after the GRB was triggered and

lasting for 2 hours). The flux and maximum energy of the afterglow emission both decrease over time, owing to deceleration of the jets – the two narrow, oppositely directed channels through which most of the explosive energy of a GRB is released. Consequently, the detection of such high-energy photons deep in the afterglow phase is also groundbreaking.

The MAGIC and HESS observatories both use an array of optical telescopes called imaging atmospheric Cherenkov telescopes (IACTs), which are designed to detect γ-rays in the very-high-energy range (roughly from 30 GeV to 100 TeV). More precisely, the IACTs detect the light (known as Cherenkov radiation) that is produced when such γ-rays hit Earth's atmosphere and produce a shower of charged particles. These facilities have been operating for more than a decade. GRBs, as the most powerful explosions in the Universe, have been one of the main observational targets, but, until now, have evaded detection. The current results are therefore a triumph for these observatories.

The discoveries are also a triumph for GRB theories. Theoretically, there are three mechanisms by which high-energy γ-rays can be produced during the afterglow phase[5]. The first is synchrotron radiation from electrons accelerated by the external shock – the shock wave that is generated when the exploded matter collides with surrounding interstellar gas. This emission component has a maximum energy that depends only on the Lorentz factor of the outflow (a parameter that denotes how fast the external shock is moving). To reach energies above 100 GeV, the Lorentz factor must be greater[5] than about 1,000, which is only marginally possible. Observations show that the Lorentz factor of GRB jets is usually a few hundred during the prompt-emission phase and decreases over time during the afterglow phase[6].

The second high-energy radiation mechanism is synchrotron radiation from protons accelerated by the GRB external shock. This emission component can, in principle, contain TeV γ-rays. However, because protons are much less efficient emitters than are electrons, the conditions for this mechanism to be dominant are rather demanding. Finally, the third mechanism is called synchrotron self-Compton[7] (SSC), whereby the same accelerated electrons that emit synchrotron photons can scatter off some of these photons, resulting in photons that have energies above 100 GeV (Fig. 1a). For typical shock-microphysics parameters inferred from afterglow modelling of other GRBs, it is expected that the SSC mechanism should usually be the main way in which high-energy γ-rays are produced[5].

One key prediction of the SSC mechanism is that there should be two 'humps' in the spectral energy distribution of the afterglow spectrum[5,7] (Fig. 1b). Such a two-hump structure has been commonly observed for high-energy jets launched from supermassive black holes known as blazars[8], and the same structure has been confidently expected for GRBs. Previous observations of high-energy afterglows of GRBs using the Large Area Telescope on board the Fermi satellite have not convincingly shown the existence of a second hump in the spectral energy distributions[9]. However, some tentative evidence has been collected from another bright burst[10–12], GRB 130427A.

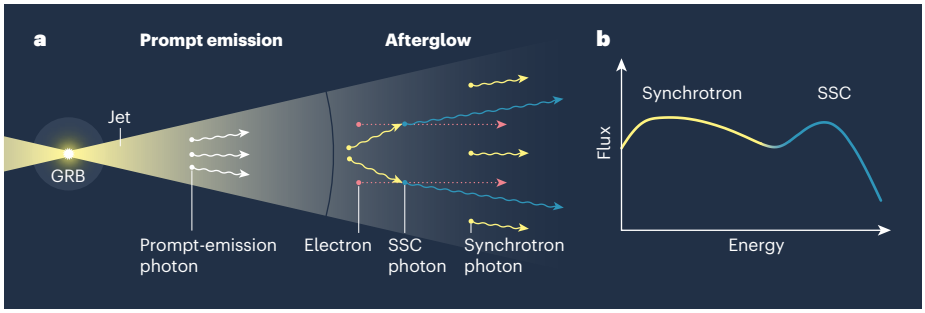


Figure 1 | Emission from a γ-ray burst. a, Three papers[2–4] report the detection of high-energy radiation from astrophysical explosions known as γ-ray bursts (GRBs). The explosive energy from a GRB is thought to be channelled into two narrow jets. Photon emission occurs in two stages: the prompt-emission phase and the afterglow phase. In the afterglow phase, low-energy photons are thought to be generated by a mechanism called synchrotron radiation. High-energy photons are thought to be mainly produced through a process dubbed synchrotron self-Compton (SSC), whereby the scattering of synchrotron photons off energetic electrons gives the photons a boost in energy[5,7]. **b**, One key prediction of the SSC mechanism is that there should be two 'humps' in the spectral energy distribution of the afterglow spectrum: one corresponding to synchrotron photons and the other to SSC photons[5,7]. Results from the three papers firmly establish the existence of such an SSC component.

The multi-wavelength observations of GRB 190114C obtained by the MAGIC Collaboration and colleagues have firmly established, for the first time, the existence of the SSC component in a GRB afterglow[3]. This conclusion has been confirmed by independent modelling from other groups[13–15]. The double-hump feature is comparatively less clear in the spectral energy distribution obtained by Abdalla *et al.* for GRB 180720B. However, in the late afterglow phase, electron synchrotron radiation cannot produce photons of energies above 100 GeV without the need to introduce exotic particle-acceleration mechanisms. As a result, the SSC mechanism is the preferred explanation for the observed spectral energy distribution[4,13].

Why did it take so long to detect a theoretically expected common spectral component? The observation of a GRB by an IACT requires that the burst is bright (to produce a sufficient number of high-energy photons) and nearby (to avoid absorption of the photons by the infrared background radiation in the Universe). Furthermore, the telescope needs to have the correct observational conditions. For instance, a particular GRB would not be detected by an IACT if the event occurred during the daytime, in poor weather or in an area of the sky that was not accessible by the telescope. Nevertheless, the breakthrough results reported in the current papers suggest that, with dedication and probably a bit of luck, a revolutionary discovery can be made.

Now that photons of energies above 100 GeV have been detected from GRBs, it is expected that such detections will become routine in the future – especially with the full operations of the available IACTs and of observatories that use other detection techniques, such as the High-Altitude Water Cherenkov Observatory in Mexico. The field will also greatly benefit from the operations of facilities such as the future international Cherenkov Telescope Array and the Large High Altitude Air Shower Observatory in Daocheng, China. As history has repeatedly shown, the opening of a new spectral window in GRB research always reveals many treasures for researchers to mine. This spectral window at the highest energies will not be any different, and could be even more rewarding.

Bing Zhang is in the Department of Physics and Astronomy, University of Nevada, Las Vegas, Nevada 89154, USA.
e-mail: zhang@physics.unlv.edu

1. Zhang, B. *The Physics of Gamma-Ray Bursts* (Cambridge Univ. Press, 2018).
2. MAGIC Collaboration. *Nature* **575**, 455–458 (2019).
3. MAGIC Collaboration *et al. Nature* **575**, 459–463 (2019).
4. Abdalla, H. *et al. Nature* **575**, 464–467 (2019).
5. Zhang, B. & Mészáros, P. *Astrophys. J.* **559**, 110–122 (2001).
6. Racusin, J. L. *et al. Astrophys. J.* **738**, 138 (2011).
7. Sari, R. & Esin, A. A. *Astrophys. J.* **548**, 787–799 (2001).
8. Ghisellini, G., Righi, C., Costamante, L. & Tavecchio, F. *Mon. Not. R. Astron. Soc.* **469**, 255–266 (2017).
9. Kumar, P. & Barniol Duran, R. *Mon. Not. R. Astron. Soc.* **400**, L75–L79 (2009).
10. Ackermann, M. *et al. Science* **343**, 42–47 (2014).
11. Liu, R.-Y., Wang, X.-Y. & Wu, X.-F. *Astrophys. J.* **773**, L20 (2013).
12. Fan, Y.-Z. *et al. Astrophys. J.* **776**, 95 (2013).
13. Wang, X.-Y., Liu, R.-Y., Zhang, H.-M., Xi, S.-Q. & Zhang, B. *Astrophys. J.* **884**, 117 (2019).
14. Derishev, E. & Piran, T. *Astrophys. J.* **880**, L27 (2019).
15. Fraija, N. *et al. Astrophys. J.* **883**, 162 (2019).

Immunotherapy

Three is a charm for an antibody to fight cancer

Alfred L. Garfall & Carl H. June

Immunotherapy approaches seek to boost immune responses against cancer. A single antibody engineered to recognize three targets shows promise, when tested in animals, in improving the ability of T cells to target cancer.

Antibodies with specificity for one target — called monoclonal antibodies — were the first cancer immunotherapy to achieve widespread clinical use. The therapeutic potency of antibodies can be amplified by engineering them to recognize two distinct molecular targets (termed antigens). These bispecific antibodies can simultaneously bind to cancer cells and immune cells called T cells, and this dual binding directs the T cell to unleash its cell-killing power towards the cancer cell. Writing in *Nature Cancer*, Wu *et al.*[1] now report the development of a trispecific antibody, one that has three targets: a cancer cell, a receptor that activates T cells, and a T-cell protein that promotes long-lasting T-cell activity against the cancer cell (Fig. 1).

The mammalian immune system generates an immense diversity of antibodies, and antibodies can also be engineered to recognize therapeutic targets. Antibodies usually recognize a single antigen, which might be part of a disease-causing agent or an abnormal version of a protein or sugar. Such monospecific antibodies against targets on cancer cells can recruit immune cells, including neutrophils, natural killer cells and macrophages, to kill or ingest the cancer cells.

Antibodies can also be engineered to block or stimulate the function of the proteins to which they bind. For example, there are regulatory receptors that inhibit T-cell function, and antibodies that have been engineered to block these receptors provide a clinical strategy known as checkpoint blockade, which boosts T-cell function. These inhibitory receptors govern T-cell exhaustion, a non-functional T-cell state that protects against autoimmunity and that can occur in the tumour microenvironment as cancers evade antitumour responses mediated by T cells. Checkpoint-blockade treatment can awaken exhausted antitumour T cells to great clinical benefit, but it also risks causing autoimmune toxicity. The antibody developed by Wu and colleagues takes a similar approach to promote T-cell activity against cancer cells. However, their method stimulates the function of receptors that positively boost T-cell function, rather than blocking the function of inhibitory receptors.

The human antibody developed by Wu *et al.* builds on bispecific-antibody technology that reconfigures the antigen-recognition domains of two different antibodies into one bispecific molecule. Bispecific antibodies usually target one antigen on the cancer cell's surface and one on a protein complex on T cells called CD3. CD3 is part of the T-cell receptor (TCR) complex. The TCR also includes antigen-recognition domains and delivers an activating signal to the T cell when an antigen binds. Engagement of CD3 by the antibody also generates an activating signal. Such a bispecific antibody therefore activates T cells, brings them into close proximity to cancer cells — irrespective of the T cell's natural antigen specificity — and redirects their killing capabilities towards the cancer cells.

This concept has proved to be clinically effective for the bispecific antibody blinatumomab, which targets CD3 and the protein CD19 on cancer cells. Blinatumomab treatment doubles the remission rate and survival among people with an advanced stage of a cancer called B-cell acute lymphoblastic leukaemia (B-ALL)[2], and it is being tested as part of the initial therapy for B-ALL, with promising early results[3].

Wu and colleagues devised a clever strategy to simultaneously boost T-cell activation and enhance the targeting of cancer cells in relation to multiple myeloma, which is a cancer of plasma cells in the blood. The authors developed a trispecific antibody that was engineered to have three antigen-binding sites, rather than two. This trispecific antibody targets CD3 plus the proteins CD38 (on cancer cells) and CD28 (on T cells). The CD38-targeting antibody daratumumab is clinically effective in treating this disease[4], and CD38 is also a potential target in other cancers, such as acute lymphoid leukaemia and acute myeloid leukaemia.

CD28 belongs to a class of protein called co-stimulatory receptors, which positively regulate T-cell activation. When a T cell recognizes its target antigen through the TCR, the extra engagement of a co-stimulatory receptor such as CD28 is needed to achieve the sustained T-cell proliferation required for

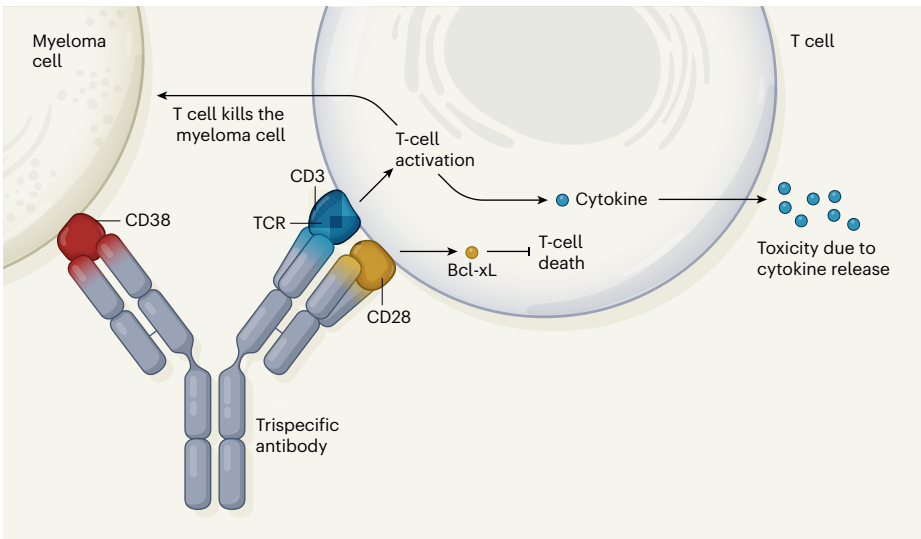


Figure 1 | An antibody that helps immune cells to target cancer cells. Wu *et al.*[1] report the development of a human antibody that is engineered to bring an immune cell called a T cell into close proximity with a type of cancer cell called a myeloma cell and to boost the T cell's anticancer response. This trispecific antibody binds three targets: the protein CD38 on a myeloma cell, and the protein CD28 and the protein complex CD3 on a T cell (the antibody's target-binding domains are shown in red, blue and yellow, respectively). CD3 is part of the T-cell receptor (TCR), which recognizes abnormal cells by binding molecules called antigens. The binding of CD3 by the antibody drives T-cell activation (without requiring antigen recognition by the TCR), which leads to the killing of the myeloma cell and the production and release of toxic cytokine molecules. Binding of CD28 by the antibody drives expression of the protein Bcl-xL. Bcl-xL blocks T-cell death, which might otherwise occur if there was prolonged TCR activation in the absence of CD28 stimulation by the antibody.

an effective immune response. In the absence of co-stimulation, activation through the TCR can lead to a state of T-cell non-responsiveness called anergy, or to the related state of exhaustion. Prolonged activation of the TCR without co-stimulation can lead the T cell to undergo a form of programmed cell death called apoptosis.

The addition of a co-stimulatory signal such as CD28 is notable because this signal has also been incorporated into another type of immunotherapy called chimaeric-antigen receptor T cell (CAR-T) therapy[5], in which a receptor is engineered to both recognize a cancer-cell antigen and include T-cell activation domains such as CD3 and CD28. The main reason for including a CD28-binding domain in the trispecific antibody is T-cell co-stimulation. However, CD28 is also frequently expressed by multiple myeloma cells, so this might increase the antibody's affinity for the myeloma cells, and thus enable it to bind to cells in which CD38 is low, absent or masked by previous daratumumab therapy.

To confirm that the CD28-binding domain augmented the trispecific antibody's activity, the authors made versions of the antibody in which different combinations of the three binding domains were mutated. They tested these versions in 'humanized' model mice, which had human T cells and human myeloma cells. A functional CD28-targeting domain boosted T-cell activation above that observed using antibodies lacking this domain. This augmented T-cell activation drove T-cell proliferation and the expression of the anti-apoptotic protein Bcl-xL in T cells, supporting the authors' hypothesis that having a co-stimulatory signal would prevent T-cell apoptosis. The presence of the CD28-targeting domain on the antibody boosted the ability of T cells to kill different myeloma cell lines *in vitro* and in the humanized mouse model, even at the lowest antibody dose tested.

The main limitation of this study is that the risk of a side effect called cytokine release syndrome (CRS), which can occur if the immune system is highly stimulated, is unknown. In CRS, the simultaneous activation of many T cells causes excessive release of signalling molecules called cytokines from cells of the immune system, which drives inflammation. CRS can occur with bispecific antibodies and with CAR-T. It typically manifests as fever, but can progress to fatal multi-organ failure in severe cases[6].

The authors report cytokine-related toxicities with their trispecific antibody when administered to monkeys by intravenous injection, but toxicity was less if it was delivered under the skin (subcutaneously) instead, leading to a more gradual exposure to the antibody. It is reassuring that the inclusion of the CD28-targeting domain did not lead to overwhelming CRS in these tests. However, a key caveat is that the amount of CD38 in monkeys is much less than in people with multiple myeloma, and the higher amount of CD38, and thus of antibody-mediated T-cell activation, would probably increase the risk of CRS in humans. But in terms of possible negative effects of the antibody on healthy non-cancerous cells, it is reassuring that only transient decreases in the number of normal white blood cells that express CD38, such as lymphocytes and myeloid cells, were observed in monkeys treated with the antibody. Another limitation of the study is that the authors did not assess whether this trispecific antibody format might trigger an immune response against the antibody and cause its rapid destruction.

"Targeting cancer using a trispecific antibody is an important conceptual advance."

Targeting cancer using a trispecific antibody is an important conceptual advance, building on previous work by this group[7] on a trispecific antibody that targets HIV. For multiple myeloma, fresh therapeutic approaches are needed, because even the most potent emerging therapies, including a CAR-T that targets an antigen called BCMA, are only temporarily effective for most people[8–10]. A trispecific antibody is a flexible platform that might offer a way to deliver precise combinations of immunomodulatory signals (for example, a co-stimulatory signal and a checkpoint blocker) specifically in the tumour microenvironment, which might be safer and more effective than the systemic administration of combinations of individual, single-specificity immunomodulatory antibodies. Such efforts to make immunotherapy more precise and potent than it is at present might be necessary to broaden the reach of immunotherapy to include the many types of cancer that have so far proved difficult to target.

Alfred L. Garfall and **Carl H. June** are in the Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.
e-mail: cjune@upenn.edu

1. Wu, L. *et al. Nature Cancer* https://doi.org/10.1038/s43018-019-0004-z (2019).
2. Kantarjian, H. *et al. N. Engl. J. Med.* **376**, 836–847 (2017).
3. Gökbuget, N. *et al. Blood* **131**, 1522–1531 (2018).
4. Lokhorst, H. M. *et al. N. Engl. J. Med.* **373**, 1207–1219 (2015).
5. June, C. H. & Sadelain, M. *N. Engl. J. Med.* **379**, 64–73 (2018).
6. Lee, D. W. *et al. Biol. Blood Marrow Transplant.* https://doi.org/10.1016/j.bbmt.2018.12.758 (2019).
7. Xu, L. *et al. Science* **358**, 85–90 (2017).
8. Raje, N. *et al. N. Engl. J. Med.* **380**, 1726–1737 (2019).
9. Cohen, A. D. *et al. J. Clin. Invest.* **130**, 2210–2221 (2019).
10. Brudno, J. N. *et al. J. Clin. Oncol.* **36**, 2267–2280 (2018).

Microbiology

Microbial clues to a liver disease

Martha R. J. Clokie

Treatment options are limited for alcoholic hepatitis, a liver disease associated with high alcohol intake. Studies in mice reveal that the microorganisms responsible for this condition can be tackled by a viral treatment. **See p.505**

In 1984, the microbiologist Barry Marshall notoriously used himself as an experimental subject for his research, and drank the contents of a flask containing the bacterium *Helicobacter pylori* as part of his efforts to demonstrate that bacteria cause stomach ulcers[1]. On page 505, Duan *et al.*[2] do not report taking such drastic action to investigate a bacterial connection to disease. Nevertheless, their careful analysis of a liver disease called alcoholic hepatitis, in studies of mice and analysis of samples from people who have the disease, also provide attention-grabbing evidence for the involvement of a suspected bacterial culprit.

Alcoholic hepatitis is a poorly understood condition related to high alcohol intake, and is difficult to treat. Previous experiments in mice have hinted that the gut-dwelling bacterium *Enterococcus faecalis* might be involved[3]. However, *E. faecalis* is usually thought of as an old friend that inhabits the guts of many animals across the evolutionary tree, from humans to nematode worms[4]. This species usually represents less than 0.1% of all the bacteria in faecal samples from healthy people[5]. However, after antibiotic treatment, bacteria of the genus *Enterococcus* increase in prevalence to become one of the most common types of microbe in the gut[6]. *E. faecalis* can infect the

https://doi.org/10.1038/d41586-019-03417-3

News & views

Microbiology

Microbial clues to a liver disease

Martha R. J. Clokie

Treatment options are limited for alcoholic hepatitis, a liver disease associated with high alcohol intake. Studies in mice reveal that the microorganisms responsible for this condition can be tackled by a viral treatment.

In 1984, the microbiologist Barry Marshall notoriously used himself as an experimental subject for his research, and drank the contents of a flask containing the bacterium *Helicobacter pylori* as part of his efforts to demonstrate that bacteria cause stomach ulcers[1]. Writing in *Nature*, Duan *et al.*[2] do not report taking such drastic action to investigate a bacterial connection to disease. Nevertheless, their careful analysis of a liver disease called alcoholic hepatitis, in studies of mice and analysis of samples from people who have the disease, also provide attention-grabbing evidence for the involvement of a suspected bacterial culprit.

Alcoholic hepatitis is a poorly understood condition related to high alcohol intake, and is difficult to treat. Previous experiments in mice have hinted that the gut-dwelling bacterium *Enterococcus faecalis* might be involved[3]. However, *E. faecalis* is usually thought of as an old friend that inhabits the guts of many animals across the evolutionary tree, from humans to nematode worms[4]. This species usually represents less than 0.1% of all the bacteria in faecal samples from healthy people[5]. However, after antibiotic treatment, bacteria of the genus *Enterococcus* increase in prevalence to become one of the most common types of microbe in the gut[6]. *E. faecalis* can infect the blood, heart, bladder and brain, and teeth that have undergone root-canal surgery[7,8].

Duan and colleagues analysed human faecal samples. They identified *E. faecalis* in the stools of about 80% of people with alcoholic hepatitis that they tested, and about 30% of the strains of *E. faecalis* present had genes that encode a toxin called cytolysin. Furthermore, people with the disease had almost 3,000 times more *E. faecalis* in their stool samples than did people who did not have alcoholic hepatitis. That isn't concrete proof that the disease is caused by this bacterium. However, the authors' data also show that the presence of cytolysin in stools correlates with mortality – 89% of the people whose faecal samples contained cytolysin died within 180 days of hospitalization, compared with only 3.8% of the people who had alcoholic hepatitis but whose stool samples lacked the toxin.

The authors next examined the connection between *E. faecalis* and liver disease in mice. The animals were colonized with strains of *E. faecalis* that either did or didn't make cytolysin, and some were then fed a high-alcohol diet, with others given an alcohol-free diet. Only the mice on the high-alcohol diet and that had been colonized with cytolysin-producing *E. faecalis* developed liver damage (Fig. 1a).

Then, using germ-free mice (which had no natural microorganisms), the authors transplanted stool samples from people with alcoholic hepatitis that contained *E. faecalis* strains in which cytolysin was either present or absent. Mice on a high-alcohol diet that were colonized with stools containing cytolysin displayed a range of signs indicating liver damage and the death of liver cells, whereas animals on such a diet and colonized with stools lacking cytolysin showed no major signs of liver damage.

To understand the disease-causing mechanisms, the authors isolated liver cells from the animals, and found that cell death in response to cytolysin exposure was dose-dependent. The response to cytolysin was the same whether or not the mice had received a high-alcohol diet. This suggests that, rather than alcohol causing alcoholic hepatitis by damaging the liver cells,

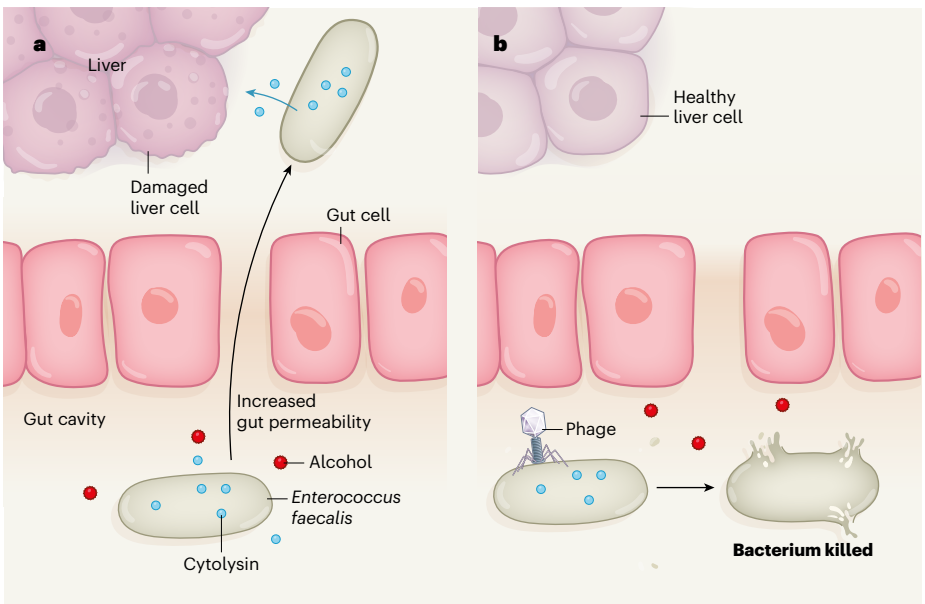


Figure 1 | Alcoholic hepatitis. Duan *et al.*[2] report studies in mice of a liver disease called alcoholic hepatitis, and analysis of faecal samples from people who have the disease. **a**, The authors report that alcoholic hepatitis is associated with the presence of a strain of the gut-dwelling bacterium *Enterococcus faecalis* that makes a toxin called cytolysin. These bacteria damage or kill liver cells, and the authors suggest that a high-alcohol diet increases gut permeability, thereby enabling the bacteria to move from the gut to the liver. **b**, To investigate possible new treatments for the disease, the authors explored the use of bacterium-targeting viruses called phages that specifically act on cytolysin-producing *E. faecalis*. When treated with these phages, *E. faecalis*-infected mice given a high-alcohol diet did not develop liver disease.

damage arises because alcohol increases the permeability of the gut lining to allow cytolysin-producing *E. faecalis* to reach the liver and cause disease symptoms (Fig. 1a).

Given the limited treatment options for alcoholic hepatitis, the authors investigated whether steps might be taken to develop a therapy that exploits bacterium-targeting viruses called bacteriophages, or phages for short. Phages have the advantage over antibiotics of being highly specific, and so avoid also killing beneficial bacteria. Furthermore, because the surface of a human cell differs substantially from that of a bacterial cell, phages aren't thought to infect animal or human cells[9].

Phages have been used to remove *Salmonella* and *Shigella* bacteria from infected human intestines for almost 100 years[10]. They have also been used to remove the disease-causing bacterium *Clostridium difficile* from artificial intestines, and from hamsters infected with this bacterium[11,12]. It has been suggested that they might one day be used in humans or animals to remodel the composition of the community of gut microorganisms (the microbiota), to produce a healthier microbiota consisting of more bacteria associated with good health and fewer associated with disease[13]. The potential of *E. faecalis*-targeting phages to tackle human diseases is already being discussed[7], and phages can kill antibiotic-resistant strains of *E. faecalis* associated with human bone and wound infections[14,15] and dental cavities[16]. Furthermore, phages are being developed for use in the food industry to remove *E. faecalis* from cheese cultures to prevent the production of toxic waste products[17].

To test whether a method could be developed to specifically remove cytolysin-producing *E. faecalis* from mice, the authors identified some phages that target these bacteria (Fig.1b) but leave other gut bacteria unaffected. Mice that received human stool samples and a high-alcohol diet and that were given *E. faecalis*-targeting phages had less liver damage than did mice given phages that killed a different bacterium not usually found in animals.

This study demonstrates the advantages of using phages in detective work to investigate the contributions of microbes to disease. The authors show that phages can be used to identify disease-causing bacterial components, and also raise the possibility that phages might offer potential treatment options. Further tests, including clinical trials, would be required to assess whether a phage approach would be useful in a human context. For example, phage treatment might help to target *E. faecalis* in the gut before a person receives a liver transplant.

In Duan and colleagues' study, the phages could treat a disease in which a causal component is a bacterium that normally resides in the gut, even though the disease site is elsewhere in the body. Although much phage research focuses on the use of these viruses to treat diseases associated with antibiotic-resistant bacteria, the work by Duan *et al.* raises the possibility of a much wider clinical role for them. There is growing evidence that gut microbes can affect the function of certain cells in the brain, and studies are ongoing to determine whether such microbes have a role in human brain diseases (see go.nature.com/2cp1kfk). Perhaps phages could become part of the next generation of targeted antimicrobial therapies for diseases that are currently difficult to treat. Indeed, there might be many diseases that we currently don't realize have a microbial component, and which could be tackled by phages.

Martha R. J. Clokie is in the Department of Genetics and Genome Biology, University of Leicester, Leicester LE1 7RH, UK.
e-mail: mrjc1@leicester.ac.uk

1. Marshall, B. J., Armstrong, J. A., McGechie, D. B. & Glancy, R. J. *Med. J. Aust.* **142**, 436–439 (1985).
2. Duan, Y. *et al. Nature* https://doi.org/10.1038/s41586-019-1742-x (2019).
3. Llorente, C. *et al. Nature Commun.* **8**, 837 (2017).
4. Van Tyne, D. & Gilmore, M. S. *Annu. Rev. Microbiol.* **68**, 337–356 (2014).
5. Lebreton, F. *et al. Cell* **169**, 849–861 (2017).
6. Ubeda, C. *et al. J. Clin. Invest.* **120**, 4332–4341 (2010).
7. Bolocan, A. S. *et al. Viruses* **11**, 366 (2019).
8. Arias, C. A. & Murray, B. E. *Nature Rev. Microbiol.* **10**, 266–278 (2012).
9. Nguyen, S. *et al. mBio* **8**, e01874–17 (2017).
10. Abedon, S. T., Kuhl, S. J., Blasdel, B. G. & Kutter, E. M. *Bacteriophage* **1**, 66-85 (2011).
11. Nale, J. Y., Chutia, M., Carr, P., Hickenbotham, P. T. & Clokie, M. R. J. *Front. Microbiol.* **7**, 1383 (2016).
12. Nale, J. Y., Redgwell, T. A., Millard, A. & Clokie, M. R. J. *Antibiotics* **7**, 13 (2018).
13. Whiteson, K. L. *mSystems* **3**, 00166-17 (2018).
14. Melo, L. D. R., Ferreira, R., Costa, A. R., Oliveira, H. & Azeredo, J. *Sci. Rep.* **9**, 6643 (2019).
15. Barros, J. *et al. Int. J. Antimicrob. Agents* **54**, 329–337 (2019).
16. Al-Zubidi, M. *et al. Infect Immun.* **87**, e00512-19 (2019).
17. del Rio, B. *et al. Front. Microbiol.* **10**, 566 (2019).

generation of targeted antimicrobial therapies for diseases that are currently difficult to treat. Indeed, there might be many diseases that we currently don't realize have a microbial component, and which could be tackled by phages.

Martha R. J. Clokie is in the Department of Genetics and Genome Biology, University of Leicester, Leicester LE1 7RH, UK.
e-mail: mrjc1@leicester.ac.uk

1. Marshall, B. J., Armstrong, J. A., McGechie, D. B. & Glancy, R. J. *Med. J. Aust.* **142**, 436–439 (1985).
2. Duan, Y. *et al. Nature* **575**, 505–511 (2019).
3. Llorente, C. *et al. Nature Commun.* **8**, 837 (2017).
4. Van Tyne, D. & Gilmore, M. S. *Annu. Rev. Microbiol.* **68**, 337–356 (2014).
5. Lebreton, F. *et al. Cell* **169**, 849–861 (2017).
6. Ubeda, C. *et al. J. Clin. Invest.* **120**, 4332–4341 (2010).
7. Bolocan, A. S. *et al. Viruses* **11**, 366 (2019).
8. Arias, C. A. & Murray, B. E. *Nature Rev. Microbiol.* **10**, 266–278 (2012).
9. Nguyen, S. *et al. mBio* **8**, e01874–17 (2017).
10. Abedon, S. T., Kuhl, S. J., Blasdel, B. G. & Kutter, E. M. *Bacteriophage* **1**, 66-85 (2011).
11. Nale, J. Y., Chutia, M., Carr, P., Hickenbotham, P. T. & Clokie, M. R. J. *Front. Microbiol.* **7**, 1383 (2016).
12. Nale, J. Y., Redgwell, T. A., Millard, A. & Clokie, M. R. J. *Antibiotics* **7**, 13 (2018).
13. Whiteson, K. L. *mSystems* **3**, 00166-17 (2018).
14. Melo, L. D. R., Ferreira, R., Costa, A. R., Oliveira, H. & Azeredo, J. *Sci. Rep.* **9**, 6643 (2019).
15. Barros, J. *et al. Int. J. Antimicrob. Agents* **54**, 329–337 (2019).
16. Al-Zubidi, M. *et al. Infect Immun.* **87**, e00512-19 (2019).
17. del Rio, B. *et al. Front. Microbiol.* **10**, 566 (2019).

This article was published online on 13 November 2019.

Engineering

Virtual and augmented reality enhanced by touch

Xiao-ming Tao

Conventional technologies for virtual and augmented reality simulate interactive experiences through visual and auditory stimuli. A technology that adds sensations of touch could find uses in areas from gaming to prosthetic feedback. **See p.473**

Human sensation includes the commonly known senses and less-recognized ones such as thirst, hunger and balance. Stimuli detected by sensory receptors are encoded into electrical signals that move along neural pathways to specific parts of the brain to be decoded into useful information. The whole process is complex. For instance, the sense of touch is a collection of several sensations, encompassing pressure, pain and temperature, and touch receptors are stimulated by a combination of mechanical, chemical and thermal energy. Until now, it has been a great challenge to incorporate sensations of touch into virtual and augmented reality. But on page 473, Yu *et al.* report a skin-integrated technology that applies pressure, vibration or motion to the user, enabling communication between the user and a machine for virtual and augmented reality (X. Yu *et al. Nature* **575**, 473–479; 2019).

The authors' technology consists of a soft, lightweight sheet of electronics that adheres to skin, and conforms to the body's shape, in a convenient, non-invasive and reversible manner (Fig. 1). The sheet contains arrays of vibratory actuators — mechanical components that convert electrical energy into vibrations. Each actuator comprises two connected parts: a coil of copper wire sealed in an acrylic base, and a permanent magnet mounted on a polymer beam. When an electric current passes through the coil, the magnet vibrates at the same frequency as that of the current.

Each actuator has a mass of only 1.4 grams and is millimetre-sized (12–18 mm in diameter and 2.5 mm thick). Given that human skin can detect submillimetre-scale touch patterns, one might question whether the actuators can be scaled down. The authors proposed a method to achieve such miniaturization and tested it by running simulations. They found that the diameter and thickness of each actuator could in future be reduced by a factor of ten and three, respectively.

A key feature of Yu and colleagues' device is that its actuator arrays are wirelessly powered and controlled. It is therefore less cumbersome than wearable platforms that require connecting wires or internal batteries. The system uses a primary antenna for power transmission, a few other antennas for controlling and driving the actuators, and an intermediate antenna to boost the power harvested from the primary antenna. Yu *et al.* found that the inclusion of the intermediate antenna increased the collected power by a factor of about three. The authors carried out simulations to confirm that their device complies with guidelines from the US Federal Communications Commission and the Federal Drug Administration regarding safe levels of radiation exposure and tissue absorption.

The distance between the power source and the platform needs to be less than about one metre, making the technology suitable for certain applications in virtual and augmented reality. Yu *et al.* describe three particular examples. In the first one, a girl touches a screen that displays a video feed of her grandmother; the grandmother senses the touch through devices on her hand and arm. In the second example, a man who has a lower-arm amputation grasps an object using a prosthetic arm that has a robotic hand; a device on his upper arm generates a pattern of sensation that reproduces the object's shape. In the third

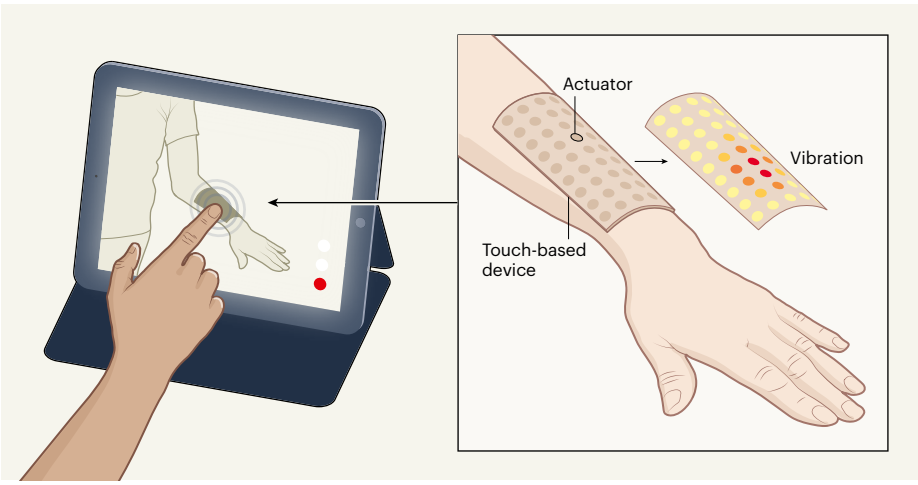


Figure 1 | Sense of virtual touch. Yu *et al.* present a device for incorporating touch-based sensations in virtual and augmented reality. The device consists of a lightweight sheet of electronics that softly laminates onto the skin. In this simple example, a touch screen displays a video feed of a person wearing the device, and a second person touches the image of the device on the screen. Mechanical components called vibrational actuators apply vibrations to the skin of the person wearing the device, providing a sense of virtual touch. The colours of the actuators represent their degree of activation from low (yellow) to high (red).

example, a person playing a combat-based video game wears several devices across their body; the devices are activated when a strike occurs on the corresponding body part of the game character.

The technology does have some drawbacks. For instance, each actuator is driven by a set current of about 5 milliamps, which is relatively high compared with that found in other consumer digital electronics. In addition, the energy lost from the other components as heat might affect actuator performance and cause warming of the skin if dissipation of the heat is not well managed. Moreover, although an optimized actuator requires only 1.75 milliwatts of power, the overall power consumption of the technology is still a key limiting factor in operating the platform sustainably and wirelessly for practical use. Miniaturization of the actuators could be a feasible way to address these issues, as the authors point out.

In 2002, many people were inspired by a smart wearable invention known as the Hug Shirt, which allows hugs to be sent over a distance with the same ease as sending a text message or chatting (see go.nature.com/32kg1oz). This technology is equipped with embedded sensors that detect and encode the strength, duration and location of the touch, together with the skin warmth and heart rate of the sender. Through wireless communication and a control circuit, these signals are decoded to control actuators that reproduce the sensation of the hug for the receiver. Both the Hug Shirt and Yu and colleagues' device suggest that the application of touch sensations in virtual and augmented reality is just beginning, and that more exciting progress can be expected in the future.

"The device is less cumbersome than wearable technologies that require connecting wires or internal batteries."

Xiao-ming Tao is at the Research Centre for Smart Wearable Technology, Institute of Textiles and Clothing, Hong Kong Polytechnic University, Kowloon, Hong Kong.
e-mail: xiao-ming.tao@polyu.edu.hk

Article

Teraelectronvolt emission from the γ-ray burst GRB 190114C

https://doi.org/10.1038/s41586-019-1750-x

Received: 10 May 2019

Accepted: 2 September 2019

Published online: 20 November 2019

MAGIC Collaboration*

Long-duration γ-ray bursts (GRBs) are the most luminous sources of electromagnetic radiation known in the Universe. They arise from outflows of plasma with velocities near the speed of light that are ejected by newly formed neutron stars or black holes (of stellar mass) at cosmological distances[1,2]. Prompt flashes of megaelectronvolt-energy γ-rays are followed by a longer-lasting afterglow emission in a wide range of energies (from radio waves to gigaelectronvolt γ-rays), which originates from synchrotron radiation generated by energetic electrons in the accompanying shock waves[3,4]. Although emission of γ-rays at even higher (teraelectronvolt) energies by other radiation mechanisms has been theoretically predicted[5–8], it has not been previously detected[7,8]. Here we report observations of teraelectronvolt emission from the γ-ray burst GRB 190114C. γ-rays were observed in the energy range 0.2–1 teraelectronvolt from about one minute after the burst (at more than 50 standard deviations in the first 20 minutes), revealing a distinct emission component of the afterglow with power comparable to that of the synchrotron component. The observed similarity in the radiated power and temporal behaviour of the teraelectronvolt and X-ray bands points to processes such as inverse Compton upscattering as the mechanism of the teraelectronvolt emission[9–11]. By contrast, processes such as synchrotron emission by ultrahigh-energy protons[10,12,13] are not favoured because of their low radiative efficiency. These results are anticipated to be a step towards a deeper understanding of the physics of GRBs and relativistic shock waves.

GRB 190114C was first identified as a long-duration GRB by the Burst Alert Telescope (BAT) onboard the Neil Gehrels Swift Observatory (Swift)[14] and the Gamma-ray Burst Monitor (GBM) instrument onboard the Fermi satellite[15] on 14 January 2019, 20:57:03 universal time (UT) (hereafter T_0). Its duration in terms of T_{90} (the time interval containing 90% of the total photon counts) was measured to be about 116 s by Fermi-GBM[15] and about 362 s by Swift-BAT[16]. Soon afterwards, reports followed on the detection of its afterglow emission at various wavebands from 1.3 GHz to 23 GeV (ref. [17]) and the measurement of its redshift[18,19], $z = 0.4245 \pm 0.0005$ (corresponding to cosmic distance). The isotropic-equivalent energy of the emission at energy of $\varepsilon = 1–10^4$ keV during T_{90} observed by Fermi-GBM was $E_{iso} \approx 3 \times 10^{53}$ erg (1 erg = 10^{-7} J), implying that GRB 190114C was fairly energetic, but not exceptionally so compared to previous events (Methods).

Triggered by the Swift-BAT alert, the Major Atmospheric Gamma Imaging Cherenkov (MAGIC) telescopes[20,21] observed GRB 190114C from $T_0 + 57$ s until $T_0 + 15{,}912$ s (Extended Data Fig. 1). γ-rays with energies above 0.2 TeV were detected with high significance from the beginning of the observations[22,23]; in the first 20 minutes of the data, the significance of the total γ-ray signal is more than 50 standard deviations (Methods, Extended Data Fig. 2).

For cosmologically distant objects such as GRBs, the observed γ-ray spectra can be substantially modified owing to attenuation by the extragalactic background light (EBL)[24]. The EBL is the diffuse background of infrared, optical and ultraviolet radiation that permeates intergalactic space, constituting the emission from all galaxies in the Universe. γ−rays can be effectively absorbed during their propagation via photon−photon pair-production interactions with low-energy photons of the EBL; this absorption is more severe for higher photon energies and higher redshifts. The γ-ray spectrum that would be observed if the EBL was absent, referred to as the intrinsic spectrum, can be inferred from the observed spectrum by 'correcting' for EBL attenuation, assuming a plausible model of the EBL[25].

Emission from GRBs occurs in two stages, which can partially overlap in time. The 'prompt' emission phase is characterized by a brief but intense flash of γ−rays, primarily at megaelectronvolt energies. It exhibits irregular variability on timescales shorter than milliseconds and lasts up to hundreds of seconds for long-duration GRBs. These γ-rays are generated in the inner regions of collimated jets of plasma, which are ejected with ultrarelativistic velocities from highly magnetized neutron stars or black holes that form following the death of massive stars[2]. The ensuing 'afterglow' phase is characterized by emission that spans a broader wavelength range and decays gradually over much longer timescales compared to the prompt emission. This originates from shock waves caused by the interaction of the jet with the ambient gas ('external shocks'). Its evolution is typified by a power-law decay

*A list of participants and their affiliations appears at the end of the paper.

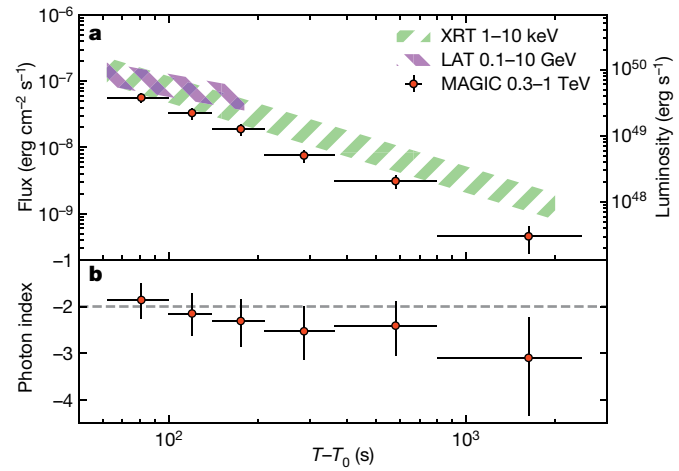


Fig. 1 | Light curves in the kiloelectronvolt, gigaelectronvolt and teraelectronvolt bands, and spectral evolution in the teraelectronvolt band for GRB 190114C. a, Light curves in units of energy flux (left axis) and apparent luminosity (right axis), for MAGIC at 0.3–1 TeV (red symbols), the Fermi Large Area Telescope (LAT) at 0.1–10 GeV (purple band) and the Swift X-ray Telescope (XRT) at 1–10 keV (green band). For the MAGIC data, the intrinsic flux is shown, corrected for EBL attenuation[25] from the observed flux. **b**, Temporal evolution of the power-law photon index, determined from time-resolved intrinsic spectra. The horizontal dashed line indicates the value −2. The errors shown in both panels are statistical only (one standard deviation).

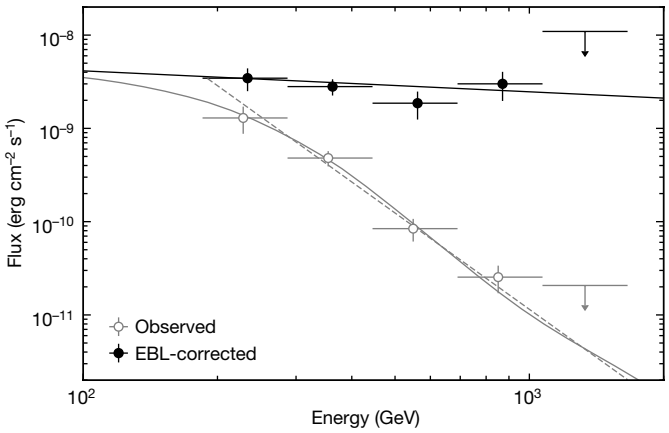


Fig. 2 | Spectrum above 0.2 TeV averaged over the period between T_0 + 62 s and T_0 + 2,454 s for GRB 190114C. Spectral-energy distributions for the spectrum observed by MAGIC (grey open circles) and the intrinsic spectrum corrected for EBL attenuation[25] (blue filled circles). The errors on the flux correspond to one standard deviation. The upper limits at 95% confidence level are shown for the first non-significant bin at high energies. Also shown is the best-fit model for the intrinsic spectrum (black curve) when assuming a power-law function. The grey solid curve for the observed spectrum is obtained by convolving this curve with the effect of EBL attenuation. The grey dashed curve is the forward-folding fit to the observed spectrum with a power-law function (Methods).

in time owing to the self-similar properties of the decelerating shock wave[3,4]. The afterglow emission of previously observed GRBs, from radio frequencies to gigaelectronvolt energies, is generally interpreted as synchrotron radiation from energetic electrons that are accelerated within magnetized plasma at the external shock[2]. Clues to whether the newly observed teraelectronvolt emission is associated with the prompt or the afterglow phase are offered by the observed light curve (flux $F(t)$ as a function of time t).

Figure 1 shows such a light curve for the EBL-corrected intrinsic flux in the energy range $\varepsilon = 0.3$–1 TeV (see also Extended Data Table 1). It is well fitted with a simple power-law function $F(t) \propto t^{\beta}$ with $\beta = -1.60 \pm 0.07$. The flux evolves from $F(t) \approx 5 \times 10^{-8}$ erg cm^{-2} s^{-1} at $t \approx T_0 + 80$ s to $F(t) \approx 6 \times 10^{-10}$ erg cm^{-2} s^{-1} at $t \gtrsim T_0 + 10^3$ s, after which it falls below the sensitivity level of the telescopes and is undetectable. There is no clear evidence for breaks or cutoffs in the light curve, nor irregular variability beyond the monotonic decay. The light curves in the kiloelectronvolt and gigaelectronvolt bands display behaviour similar to the teraelectronvolt band, with a somewhat shallower decay slope for the gigaelectronvolt band (Fig. 1). These properties indicate that most of the observed emission is associated with the afterglow phase, rather than the prompt phase, which typically shows irregular variability. We note that although the measured T_{90} is as long as about 360 s, the kiloelectronvolt–megaelectronvolt emission does not exhibit clear temporal or spectral evidence for a prompt component after about $T_0 + 25$ s (ref. [26]; Methods). Nevertheless, a sub-dominant contribution to the teraelectronvolt emission from a prompt component at later times cannot be excluded. The flux initially observed at $t \approx T_0 + 80$ s corresponds to an apparent isotropic-equivalent luminosity of $L_{\rm iso} \approx 3 \times 10^{49}$ erg s^{-1} at $\varepsilon = 0.3$–1 TeV, making this the most luminous source known at these energies.

The power radiated in the teraelectronvolt band is comparable, within a factor of about 2, to that in the soft-X-ray and gigaelectronvolt bands during the periods when simultaneous teraelectronvolt–kiloelectronvolt or teraelectronvolt–gigaelectronvolt data are available (Fig. 1). The isotropic-equivalent energy radiated at $\varepsilon = 0.3$–1 TeV, integrated over the time period between $T_0 + 62$ s and $T_0 + 2{,}454$ s, is $E_{0.3\text{–}1\rm TeV} \approx 4 \times 10^{51}$ erg. This is a lower limit to the total teraelectronvolt-band output, as it does not account for data before $T_0 + 62$ s or potential emission at $\varepsilon > 1$ TeV. From the megaelectronvolt–gigaelectronvolt data, the power-law decay phase is inferred to start at about $T_0 + 6$ s (refs. [26,27]). Assuming that the MAGIC light curve evolved as $F(t) \propto t^{-1.60}$ after that time, the teraelectronvolt-band energy integrated between $T_0 + 6$ s and $T_0 + 2{,}454$ s is $E_{0.3\text{–}1\rm TeV} \approx 2 \times 10^{52}$ erg. This would be about 10% of the $E_{\rm iso}$ value measured by Fermi-GBM at $\varepsilon = 1$–10^4 keV.

Figure 1 also shows the time evolution of the intrinsic spectral photon index $\alpha_{\rm int}$, determined by fitting the EBL-corrected, time-dependent differential photon spectrum with the power-law function $\mathrm{d}F/\mathrm{d}\varepsilon \propto \varepsilon^{\alpha_{\rm int}}$. Considering the statistical and systematic errors (Methods), there is no significant evidence for spectral variability. Throughout the observations, the data are consistent with $\alpha_{\rm int} \approx -2$, indicating that the radiated power is nearly equally distributed in ε over this band.

Figure 2 presents both the observed and the EBL-corrected intrinsic spectra above 0.2 TeV, averaged over ($T_0 + 62$ s, $T_0 + 2{,}454$ s). The observed spectrum can be fitted in the energy range 0.2–1 TeV with a simple power law with photon index $\alpha_{\rm obs} = -5.43 \pm 0.22$ (statistical error only), one of the steepest spectra ever observed for a γ-ray source. It is remarkable that photons are observed at $\varepsilon \approx 1$ TeV (Extended Data Table 2), despite the severe EBL attenuation expected at these energies (by a factor of about 300, according to plausible EBL models; see Methods). Assuming a particular EBL model[25], the intrinsic spectrum is well described as a power law with $\alpha_{\rm int} = -2.22^{+0.23}_{-0.25}$ (statistical error only), extending beyond 1 TeV at 95% confidence level with no evidence for a spectral break or cutoff (Methods). Adopting other EBL models leads to only small differences in $\alpha_{\rm int}$, which are within the uncertainties (Methods). Consistency with $\alpha_{\rm int} \approx -2$ implies a roughly equal power radiated over 0.2–1 TeV and possibly beyond, strengthening the inference that there is substantial energy output at teraelectronvolt energies.

Much of the observed emission up to gigaelectronvolt energies for GRB 190114C is probably afterglow synchrotron emission from electrons, similar to that of many previous GRBs[2,28]. The teraelectronvolt emission observed here is also plausibly associated with the afterglow. However, it cannot be a simple spectral extension of the electron synchrotron emission. The maximum energy of the emitting electrons is determined by the balance between their energy losses, which are

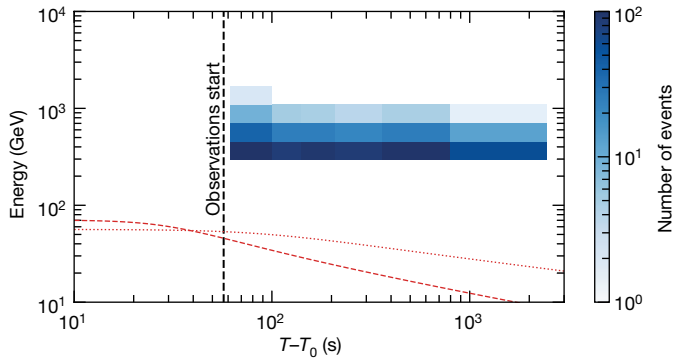


Fig. 3 | Distribution of the number of teraelectronvolt-band γ-rays in time and energy for GRB 190114C. The number of events in each bin of energy and time are colour-coded (Methods). The vertical line indicates the beginning of the data acquisition. The curves show the expected maximum photon energy $\varepsilon_{\mathrm{syn,max}}$ of electron synchrotron radiation in the standard afterglow theory for two extreme cases giving high values of $\varepsilon_{\mathrm{syn,max}}$. The dotted curve corresponds to an isotropic-equivalent blast-wave kinetic energy of $E_{\mathrm{k,aft}} = 3\times10^{55}$ erg and a homogeneous external medium with density $n = 0.01\,\mathrm{cm}^{-3}$; the dashed curve corresponds to $E_{\mathrm{k,aft}} = 3\times10^{55}$ erg and an external medium describing a progenitor stellar wind with a density profile of $n(R) = AR^{-2}$ as a function of radius R, where $A = 3\times10^{33}\,\mathrm{cm}^{-1}$ (Methods).

dominated by synchrotron radiation, and their acceleration. The timescale of the latter should not be much shorter than that of their gyration around the magnetic field at the external shock. The energy of afterglow synchrotron photons is then limited to a maximum value, the so-called synchrotron burnoff limit[29,30] of $\varepsilon_{\mathrm{syn,max}} \approx 100(\Gamma_{\mathrm{b}}/1{,}000)$ GeV, which depends only on the bulk Lorentz factor Γ_{b}. The latter is unlikely to considerably exceed $\Gamma_{\mathrm{b}} \approx 1{,}000$ (Methods). Figure 3 compares the observed photon energies with expectations of $\varepsilon_{\mathrm{syn,max}}$ under different assumptions. Although a few γ-rays with energy approaching $\varepsilon_{\mathrm{syn,max}}$ have been previously detected from a GRB by Fermi[30], the evidence for a separate spectral component was not conclusive, given the uncertainties in Γ_{b}, the electron acceleration rate and the spatial structure of the emitting region[31]. Here, even the lowest-energy photons detected by MAGIC are considerably above $\varepsilon_{\mathrm{syn,max}}$ and extend beyond 1 TeV at 95% confidence level (Methods). Thus, this observation provides the first unequivocal evidence for a new emission component beyond synchrotron emission in the afterglow of a GRB. Moreover, this component is energetically important, with a power nearly comparable to that of the synchrotron component observed contemporaneously.

Comparing with previous MAGIC observations of GRBs, the fact that GRB 190114C was the first to be clearly detected may be due to a favourable combination of its low redshift and suitable observing conditions, rather than its intrinsic properties being exceptional (Methods), although firm conclusions cannot yet be drawn with only one positive detection. The capability of the telescopes to react fast and operate during moonlight conditions was crucial in achieving this detection.

The discovery of an energetically important emission component beyond electron synchrotron emission that may be common in GRB afterglows offers important new insight into the physics of GRBs. The similarity of the radiated power and temporal decay slopes in the teraelectronvolt and X-ray bands suggests that this component is intimately related to the electron synchrotron emission. Promising mechanisms for the teraelectronvolt emission are 'leptonic' processes in the afterglow such as inverse Compton radiation, in which the electrons in the external shock Compton-scatter ambient low-energy photons to higher energies[9–11]. On the other hand, 'hadronic' processes induced by ultrahigh-energy protons in the external shock[10,12,13] may also be viable if the acceleration of electrons and protons occurs in a correlated manner. However, such processes typically have low radiative efficiency, and are not favoured as the origin of the luminous teraelectronvolt emission observed in GRB 190114C for cases such as proton synchrotron emission (Methods). Continuing efforts with existing and future γ-ray telescopes will test these expectations and provide further insight into the physics of GRBs and related issues.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1750-x.

1. Gehrels, N. & Mészáros, P. Gamma-ray bursts. *Science* **337**, 932–936 (2012).
2. Kumar, P. & Zhang, B. The physics of gamma-ray bursts & relativistic jets. *Phys. Rep.* **561**, 1–109 (2015).
3. Mészáros, P. Theories of gamma-ray bursts. *Annu. Rev. Astron. Astrophys.* **40**, 137–169 (2002).
4. Piran, T. The physics of gamma-ray bursts. *Rev. Mod. Phys.* **76**, 1143–1210 (2005).
5. Mészáros, P., Razzaque, S. & Zhang, B. GeV–TeV emission from γ-ray bursts. *New Astron. Rev.* **48**, 445–451 (2004).
6. Fan, Y.-Z. & Piran, T. High-energy γ-ray emission from gamma-ray bursts – before GLAST. *Front. Phys. China* **3**, 306–330 (2008).
7. Inoue, S. et al. Gamma-ray burst science in the era of the Cherenkov Telescope Array. *Astropart. Phys.* **43**, 252–275 (2013).
8. Nava, L. High-energy emission from gamma-ray bursts. *Int. J. Mod. Phys. D* **27**, 1842003 (2018).
9. Mészáros, P., Rees, M. J. & Papathanassiou, H. Spectral properties of blast-wave models of gamma-ray burst sources. *Astrophys. J.* **432**, 181–193 (1994).
10. Zhang, B. & Mészáros, P. High-energy spectral components in gamma-ray burst afterglows. *Astrophys. J.* **559**, 110–122 (2001).
11. Beniamini, P., Nava, L., Duran, R. B. & Piran, T. Energies of GRB blast waves and prompt efficiencies as implied by modelling of X-ray and GeV afterglows. *Mon. Not. R. Astron. Soc.* **454**, 1073–1085 (2015).
12. Vietri, M. GeV photons from ultrahigh energy cosmic rays accelerated in gamma-ray bursts. *Phys. Rev. Lett.* **78**, 4328–4331 (1997).
13. Böttcher, M. & Dermer, C. D. High-energy gamma rays from ultra-high-energy cosmic-ray protons in gamma-ray bursts. *Astrophys. J. Lett.* **499**, 131–134 (1998).
14. Gropp, J. D. GRB 190114C: Swift detection of a very bright burst with a bright optical counterpart. *GCN Circulars* 23688 https://gcn.gsfc.nasa.gov/gcn3/23688.gcn3 (2019).
15. Hamburg, R. GRB 190114C: Fermi GBM detection. *GCN Circulars* 23707 https://gcn.gsfc.nasa.gov/gcn3/23707.gcn3 (2019).
16. Krimm, H. A. et al. GRB 190114C: Swift-BAT refined analysis. *GCN Circulars* 23724 https://gcn.gsfc.nasa.gov/gcn3/23724.gcn3 (2019).
17. MAGIC Collaboration et al. Observation of inverse Compton emission from a long γ-ray burst. *Nature* https://doi.org/10.1038/s41586-019-1754-6 (2019).
18. Selsing, J. GRB 190114C: NOT optical counterpart and redshift. *GCN Circulars* 23695 https://gcn.gsfc.nasa.gov/gcn3/23695.gcn3 (2019).
19. Castro-Tirado, A. GRB 190114C: refined redshift by the 10.4m GTC. *GCN Circulars* 23708 https://gcn.gsfc.nasa.gov/gcn3/23708.gcn3 (2019).
20. Aleksić, J. et al. The major upgrade of the MAGIC telescopes, part I: the hardware improvements and the commissioning of the system. *Astropart. Phys.* **72**, 61–75 (2016).
21. Aleksić, J. et al. The major upgrade of the MAGIC telescopes, part II: a performance study using observations of the Crab Nebula. *Astropart. Phys.* **72**, 76–94 (2016).
22. Mirzoyan, R. First time detection of a GRB at sub-TeV energies; MAGIC detects the GRB 190114C. *The Astronomer's Telegram* 12390 http://www.astronomerstelegram.org/?read=12390 (2019).
23. Mirzoyan, R. et al. MAGIC detects the GRB 190114C in the TeV energy domain. *GCN Circulars* 23701 https://gcn.gsfc.nasa.gov/gcn3/23701.gcn3 (2019).
24. Dwek, E. & Krennrich, F. The extragalactic background light and the gamma-ray opacity of the universe. *Astropart. Phys.* **43**, 112–133 (2013).
25. Domínguez, A. et al. Extragalactic background light inferred from AEGIS galaxy-SED-type fractions. *Mon. Not. R. Astron. Soc.* **410**, 2556–2578 (2011).
26. Ravasio, M. E. et al. GRB 190114C: from prompt to afterglow? *Astron. Astrophys.* **626**, A12 (2019).
27. Wang, X.-Y., Liu, R.-Y., Zhang, H.-M., Xi, S.-Q. & Zhang, B. Synchrotron self-Compton emission from afterglow shocks as the origin of the sub-TeV emission in GRB 180720B and GRB 190114C. *Astrophys. J.* (in the press).
28. Ackermann, M. et al. The first Fermi-LAT gamma-ray burst catalog. *Astrophys. J. Suppl. Ser.* **209**, 11 (2013).
29. Piran, T. & Nakar, E. On the external shock synchrotron model for gamma-ray bursts' GeV emission. *Astrophys. J. Lett.* **718**, 63–67 (2010).
30. Ackermann, M. et al. Fermi-LAT observations of the gamma-ray burst GRB 130427A. *Science* **343**, 42–47 (2014).
31. Kouveliotou, C. et al. NuSTAR observations of GRB 130427A establish a single component synchrotron afterglow origin for the late optical to multi-GeV emission. *Astrophys. J. Lett.* **779**, L1 (2013).

MAGIC Collaboration*

V. A. Acciari[1], S. Ansoldi[2,3], L. A. Antonelli[4], A. Arbet Engels[5], D. Baack[6], A. Babić[7], B. Banerjee[8], U. Barres de Almeida[9], J. A. Barrio[10], J. Becerra González[1], W. Bednarek[11], L. Bellizzi[12], E. Bernardini[13,14], A. Berti[15], J. Besenrieder[16], W. Bhattacharyya[13], C. Bigongiari[4], A. Biland[5], O. Blanch[17], G. Bonnoli[12], Ž. Bošnjak[7], G. Busetto[14], A. Carosi[4,38], R. Carosi[18], G. Ceribella[16], Y. Chai[16], A. Chilingaryan[19], S. Cikota[7], S. M. Colak[17], U. Colin[16], E. Colombo[1], J. L. Contreras[10], J. Cortina[20], S. Covino[4], G. D'Amico[16], V. D'Elia[4], P. Da Vela[18], F. Dazzi[4], A. De Angelis[14], B. De Lotto[2], M. Delfino[17,21], J. Delgado[17,21], D. Depaoli[15], F. Di Pierro[15], L. Di Venere[15], E. Do Souto Espiñeira[17], D. Dominis Prester[22], A. Donini[2], D. Dorner[23], M. Doro[14], D. Elsaesser[6], V. Fallah Ramazani[24], A. Fattorini[6], A. Fernández-Barral[14], G. Ferrara[4], D. Fidalgo[10], L. Foffano[14], M. V. Fonseca[10], L. Font[25], C. Fruck[16], S. Fukami[26], S. Gallozzi[4], R. J. García López[1], M. Garczarczyk[13], S. Gasparyan[19], M. Gaug[25], N. Giglietto[15], F. Giordano[15], N. Godinović[27], D. Green[16], D. Guberman[17], D. Hadasch[26], A. Hahn[16], J. Herrera[1], J. Hoang[10], D. Hrupec[28], M. Hütten[16], T. Inada[26], S. Inoue[29], K. Ishio[16], Y. Iwamura[26], L. Jouvin[17], D. Kerszberg[17], H. Kubo[3], J. Kushida[30], A. Lamastra[4], D. Lelas[27], F. Leone[4], E. Lindfors[24], S. Lombardi[4], F. Longo[2,31,32], M. López[10], R. López-Coto[14], A. López-Oramas[1], S. Loporchio[15], B. Machado de Oliveira Fraga[9], C. Maggio[25], P. Majumdar[8], M. Makariev[33], M. Mallamaci[14], G. Maneva[33], M. Manganaro[22], K. Mannheim[23], L. Maraschi[4], M. Mariotti[14], M. Martínez[17], S. Masuda[3], D. Mazin[16,26], S. Mićanović[22], D. Miceli[2], M. Minev[33], J. M. Miranda[12], R. Mirzoyan[16], E. Molina[34], A. Moralejo[17], D. Morcuende[10], V. Moreno[25], E. Moretti[17], P. Munar-Adrover[25], V. Neustroev[35], C. Nigro[13], K. Nilsson[24], D. Ninci[17], K. Nishijima[30], K. Noda[26], L. Nogués[17], M. Nöthe[6], S. Nozaki[3], S. Paiano[14], J. Palacio[17], M. Palatiello[2], D. Paneque[16], R. Paoletti[12], J. M. Paredes[34], P. Peñil[10], M. Peresano[2], M. Persic[2], P. G. Prada Moroni[18], E. Prandini[14], I. Puljak[27], W. Rhode[6], M. Ribó[34], J. Rico[17], C. Righi[4], A. Rugliancich[18], L. Saha[10], N. Sahakyan[19], T. Saito[26], S. Sakurai[26], K. Satalecka[13], K. Schmidt[6], T. Schweizer[16], J. Sitarek[11], I. Šnidarić[36], D. Sobczynska[11], A. Somero[1], A. Stamerra[4], D. Strom[16], M. Strzys[16], Y. Suda[16], T. Surić[36], M. Takahashi[26], F. Tavecchio[4], P. Temnikov[33], T. Terzić[22,36], M. Teshima[16,26], N. Torres-Albà[34], L. Tosti[15], S. Tsujimoto[30], V. Vagelli[15], J. van Scherpenberg[16], G. Vanzo[1], M. Vazquez Acosta[1], C. F. Vigorito[15], V. Vitale[15], I. Vovk[16], M. Will[16], D. Zarić[27] & L. Nava[4,32,37]

[1]Instituto de Astrofísica de Canarias and Departamento Astrofísica, Universidad de La Laguna, La Laguna, Spain. [2]Università di Udine and INFN Trieste, Udine, Italy. [3]Japanese MAGIC Consortium, Department of Physics, Kyoto University, Kyoto, Japan. [4]National Institute for Astrophysics (INAF), Rome, Italy. [5]ETH Zurich, Zurich, Switzerland. [6]Technische Universität Dortmund, Dortmund, Germany. [7]Croatian Consortium, University of Zagreb – FER, Zagreb, Croatia. [8]Saha Institute of Nuclear Physics, HBNI, Kolkata, India. [9]Centro Brasileiro de Pesquisas Físicas (CBPF), Rio de Janeiro, Brazil. [10]IPARCOS Institute and EMFTEL Department, Universidad Complutense de Madrid, Madrid, Spain. [11]University of Łódź, Department of Astrophysics, Łódź, Poland. [12]Università di Siena and INFN Pisa, Siena, Italy. [13]Deutsches Elektronen-Synchrotron (DESY), Zeuthen, Germany. [14]Università di Padova and INFN, Padua, Italy. [15]Istituto Nazionale Fisica Nucleare (INFN), Frascati, Italy. [16]Max-Planck-Institut für Physik, Munich, Germany. [17]Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [18]Università di Pisa and INFN Pisa, Pisa, Italy. [19]The Armenian Consortium, ICRANet-Armenia at NAS RA, A. Alikhanyan National Laboratory, Yerevan, Armenia. [20]Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas, Madrid, Spain. [21]Port d'Informació Científica (PIC), Barcelona, Spain. [22]Croatian Consortium, Department of Physics, University of Rijeka, Rijeka, Croatia. [23]Universität Würzburg, Würzburg, Germany. [24]Finnish MAGIC Consortium, Finnish Centre of Astronomy with ESO (FINCA), University of Turku, Turku, Finland. [25]Departament de Física and CERES-IEEC, Universitat Autònoma de Barcelona, Bellaterra, Spain. [26]Japanese MAGIC Consortium, ICRR, The University of Tokyo, Kashiwa, Japan. [27]Croatian Consortium, University of Split – FESB, Split, Croatia. [28]Croatian Consortium, Josip Juraj Strossmayer University of Osijek, Osijek, Croatia. [29]Japanese MAGIC Consortium, RIKEN, Wako, Japan. [30]Japanese MAGIC Consortium, Tokai University, Hiratsuka, Japan. [31]Dipartimento di Fisica, Università di Trieste, Trieste, Italy. [32]Institute for Fundamental Physics of the Universe (IFPU), Trieste, Italy. [33]Institute for Nuclear Research and Nuclear Energy, Bulgarian Academy of Sciences, Sofia, Bulgaria. [34]Universitat de Barcelona, ICCUB, IEEC-UB, Barcelona, Spain. [35]Finnish MAGIC Consortium, Astronomy Research Unit, University of Oulu, Oulu, Finland. [36]Croatian Consortium, Rudjer Boskovic Institute, Zagreb, Croatia. [37]Istituto Nazionale Fisica Nucleare (INFN), Trieste, Italy. [38]Present address: Laboratoire d'Annecy de Physique des Particules, Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, LAPP, Annecy, France. *e-mail: contact.magic@mpp.mpg.de

Methods

General properties of GRB 190114C

GRB 190114C was first identified by the Swift-BAT[14] and Fermi-GBM[15] instruments on 14 January 2019, 20:57:03 UT. Subsequently, it was also detected by several other space-based instruments, including Fermi-LAT, INTEGRAL/SPI-ACS, AGILE/MCAL, Insight/HXMT and Konus-Wind[17,26]. Its redshift was reported as $z = 0.4245 \pm 0.0005$ by the Nordic Optical Telescope[18] and confirmed by Gran Telescopio Canarias[19]. The measured duration of $T_{90} \approx 116$ s by Fermi-GBM and $T_{90} \approx 362$ s by Swift-BAT[16] puts GRB 190114C unambiguously in the long-duration subclass of GRBs[1]. The fluence and peak photon flux of the emission at 10–1,000 keV during T_{90} measured by Fermi-GBM are $(3.990 \pm 0.008) \times 10^{-4}$ erg cm^{-2} and (246.86 ± 0.86) cm^{-2} s^{-1} (ref. [15]). The corresponding isotropic equivalent energy and luminosity at 1–10^4 keV are $E_{iso} \approx 3 \times 10^{53}$ erg and $L_{iso} \approx 1 \times 10^{53}$ erg s^{-1}, respectively[26]. These values are consistent with the known correlations between the spectral peak energy ε_{peak} and E_{iso} (ref. [32]) and between ε_{peak} and L_{iso} (ref. [33]) for GRBs. The light curve of the kiloelectronvolt–megaelectronvolt emission exhibits two prominent emission episodes with irregular multi-peaked structure at $t \approx 0$–5 s and $t \approx 15$–25 s (Extended Data Fig. 1). The spectra for these episodes are typical of GRB prompt emission[26]. On the other hand, at $t \approx 15$–25 s and $t \geq 25$ s, the temporal and spectral properties of the kiloelectronvolt–megaelectronvolt emission are consistent with an afterglow component, indicating a considerable overlap in time between the prompt and afterglow phases. Indeed, from a joint spectral and temporal analysis of the Fermi-GBM and Fermi-LAT data, the onset of the afterglow for GRB 190114C was estimated to occur at $t \approx 6$ s, much earlier than T_{90} (ref. [26]).

The event is fairly energetic but not exceptionally so, with E_{iso} lying in the highest ~30% of its known distribution[34]. No neutrinos were detected by the IceCube Observatory in the energy range 100 TeV to 10 PeV, under non-optimal observing conditions[35].

MAGIC telescopes and automatic alert system

The MAGIC telescopes comprise two 17-m diameter imaging atmospheric Cherenkov telescopes (IACTs; MAGIC-I and MAGIC-II) operating in stereoscopic mode, located at the Roque de los Muchachos Observatory in La Palma, Canary Islands, Spain[20,21]. By imaging Cherenkov light from extended air shower events, the telescopes can detect γ-rays above an energy threshold of 30 GeV, depending on the observing mode and conditions, with a field of view of ~10 square degrees.

Observing GRBs with IACTs such as those of MAGIC warrants a dedicated strategy. Because IACTs have a low probability of discovering GRBs serendipitously in their relatively small field of view, they rely on external alerts provided by satellite instruments with larger fields of view to trigger follow-up observations. Since their inception, the MAGIC telescopes were designed to perform fast follow-up observations of GRBs. By virtue of their light-weight reinforced-carbon-fibre structure and high repositioning speed, they can respond quickly to GRB alerts received via the Gamma-ray Coordinates Network (GCN; https://gcn.gsfc.nasa.gov)[36]. After various updates to the entire system over the years[20,21], the telescopes can currently slew to a target with a repositioning speed of 7° s^{-1}. To achieve the fastest possible response to GRB alerts, an automatic alert system (AAS) has been developed, which is a multi-threaded programme that performs different tasks, such as connecting to the GCN servers, receiving GCN notices that contain the sky coordinates of the GRB and sending commands to the Central Control (CC) software of the MAGIC telescopes. This also includes a check of the visibility of the new target according to predefined criteria. A priority list has been set up for cases in which several different types of alerts are received simultaneously. Moreover, if there are multiple alerts for the same GRB, the AAS selects the one with the best localization.

If an alert is tagged as observable by the AAS, the telescopes automatically repoint to the new sky position. An automatic procedure, implemented in 2013, prepares the subsystems for data taking during the telescope slewing[37,38]: data taken previously are saved, relevant trigger tables are loaded, appropriate electronics thresholds are set and the mirror segments are suitably adjusted by the Automatic Mirror Control hardware. While moving, the telescopes calibrate the imaging cameras. The data acquisition system continues taking data while it receives information about the target from the CC software. The presence of a trigger limiter set to 1 kHz prevents high rates and the saturation of the data acquisition system. When the repositioning has finished, the target is tracked in wobble mode, which is the standard observing mode for MAGIC[39]. The fastest so far GRB follow-up was achieved for GRB 160821B, when the data taking started only 24 s after the GRB.

MAGIC observations of GRB 190114C

On the night of 14 January 2019, at 20:57:25 UT (T_0 + 22 s), Swift-BAT distributed an alert reporting the first estimated coordinates of GRB 190114C (right ascension, +03 h 38 min 02 s; declination, −26 d 56 min 18 s). The AAS validated it as observable and triggered the automatic repointing procedure, and the telescopes began slewing in fast mode from their position before the alert. The MAGIC-I and MAGIC-II telescopes were on target and began tracking GRB 190114C at 20:57:52.858 UT and 20:57:53.260 UT (T_0 + 50 s), respectively, starting from a zenith angle of 55.8° and an azimuth angle of 175.1° in local coordinates. After starting the slewing, the telescopes reached the target position in approximately 27 s, moving by 42.82° in zenith and 177.5° in azimuth. At the end of the slewing, the cameras on the telescopes oscillated for a short time. Subsequently, we performed dedicated tests that reproduced the movement of the telescopes. We verified that the duration of the oscillations was less than 10 s after the start of the tracking, and their amplitude was less than 0.6′ when data taking began. Data acquisition started at 20:58:00 (T_0 + 57 s) and the data acquisition system was operating stably from 20:58:05 (T_0 + 62 s), as denoted in Extended Data Fig. 1.

Observations were performed in the presence of moonlight, implying a relatively high night sky background (NSB), approximately 6 times the level for dark observations (moonless nights with good weather conditions)[40]. Data taking for GRB 190114C stopped on 15 January 2019, 01:22:15 UT, when the target reached a zenith angle of 81.14° and an azimuth angle of 232.6°. The total exposure time for GRB 190114C was 4.12 h.

MAGIC data analysis for GRB 190114C

Data collected from GRB 190114C were analysed using the standard MAGIC analysis software[21] and with the analysis chain tuned for data taken under moonlight conditions[40]. No detailed information on the atmospheric transmission was available because the LIDAR facility[41] was not operating during the night of the observation. Therefore, the quality of the data was assessed by checking other auxiliary weather-monitoring devices, as well as the value and stability of the data acquisition rates.

A dedicated set of Monte Carlo simulation γ-ray data was produced for the analysis, matching the trigger settings (discriminator thresholds), the zenith–azimuth distribution and the NSB level of the GRB 190114C observations. The final dataset comprises events starting from 20:58:05 UT. Owing to the higher NSB compared to standard analysis, a higher level of image cleaning was applied to both the measured and the Monte Carlo data, and a higher cut on the integrated charge of the event image, set to 80 photoelectrons, was used for evaluating photon fluxes[40]. The significance of the γ-ray signal was computed using the Li & Ma method[42].

The spectra in Fig. 2 were derived by assuming a simple power-law function for the intrinsic spectrum

$$\frac{\mathrm{d}F}{\mathrm{d}\varepsilon} = f_0 \times \left(\frac{\varepsilon}{\varepsilon_0}\right)^{-\alpha}$$

with the forward-folding method to derive the best-fit parameters and the Schmelling unfolding prescription for the spectral points[43], starting from the observed spectrum and correcting for EBL attenuation with the model of Dominguez et al.[25]. The best-fit values are $\alpha_{\rm int} = -2.22^{+0.23}_{-0.25}(\text{statistical})^{+0.21}_{-0.26}(\text{systematic})$ and $f_{0,\rm int} = [8.45^{+0.68}_{-0.65}(\text{statistical})^{+4.42}_{-3.97}(\text{systematic})] \times 10^{-9}\ \text{TeV}^{-1}\,\text{cm}^{-2}\,\text{s}^{-1}$ at 0.46 TeV. We note that owing to the soft spectrum of the source, the systematic errors reported here are larger than those given in ref. [21].

The absolute energy scale for MAGIC measurements is systematically affected by the imperfect knowledge of different aspects, such as the atmospheric transmission, the mirror reflectance and the properties of photomultipliers. A dedicated study[21] identified the light-scale matching of measured and Monte Carlo data as the most important contribution to the systematic errors on the absolute energy scale. A miscalibration of the Monte Carlo energy scale can lead to mis-reconstruction of the spectrum that affects both the flux and the spectral shape, especially at the lowest energies. These studies demonstrated that the reconstructed spectra for MAGIC are affected by a systematic error due to the variation of the light scale by less than ±15%. In the case of moonlight observations, additional systematic effects on the flux arise from mismatches between Monte Carlo and measured data, in particular of the trigger discriminator thresholds and of the higher noise in the photomultipliers. Dedicated studies for moonlight observations[40] reveal that these errors affect only the overall flux (and not the spectral index) and depend on the NSB level. The contribution to the systematic error from the moonlight observations is minor compared to that due to the light-scale variations. Moreover, in the case of GRB 190114C, the influence of moonlight conditions on the overall systematic errors is mitigated by the improved data–Monte Carlo agreement achieved by simulating the recorded trigger discriminator thresholds and NSB during the GRB 190114C observation. For the analysis of the GRB 190114C data, we reproduced the effect of the light-scale variations on the spectra to derive the systematic errors on the energy flux and the errors on the photon index reported in Extended Data Table 1. The light-scale modifications were applied to the spectra before their deconvolution with EBL attenuation, which ultimately affects the low- and high-energy ends of the spectra in different ways. The fit to the obtained curves was performed in the same manner as the nominal case. Finally, the systematic errors were obtained from the difference of the parameter values computed for the nominal case and for the cases of light-scale variations by ±15%.

An additional systematic effect originates from uncertainties in existing EBL models. To quantify the corresponding systematic errors on the derived photon indices, the observed spectra were corrected by adopting several EBL models[44–46] for the redshift of this GRB. The results can be found in Extended Data Table 4. The spectral indices inferred using different EBL models differ less than their statistical uncertainties (one standard deviation). Taking as reference the EBL model of Dominguez et al.[25], the spectral index for the time-integrated spectrum has an additional systematic error due to uncertainties in the EBL such that $\alpha_{\rm int} = -2.22^{+0.23}_{-0.25}(\text{statistical})^{+0.21}_{-0.26}(\text{systematic})^{+0.07}_{-0.17}(\text{systematic}_{\rm EBL})$. The observed spectrum in the 0.2–1.0 TeV energy range can be roughly described by a power law with photon index $\alpha_{\rm obs} = -5.43 \pm 0.22$ (statistical) and flux normalization $f_{0,\rm obs} = [4.09 \pm 0.34\ (\text{statistical})] \times 10^{-10}\ \text{TeV}^{-1}\,\text{cm}^{-2}\,\text{s}^{-1}$ at 0.475 TeV.

The upper limit for the first non-significant energy bin in the observed spectrum shown in Fig. 2 is calculated from a likelihood ratio test between two models. The first, baseline, model considers only background events and spillover events from lower energy. The second model additionally assumes that the spectrum extends to higher energy as an unbroken power law, with the flux normalization as a free parameter. Given the low event statistics in the higher-energy bins, the validity of the upper limit was checked by performing 10,000 Monte Carlo simulations of the likelihood ratio test. The test statistic distribution derived from this toy simulation was then used to determine the upper limit on the flux at 95% confidence level. The corresponding upper limit for the intrinsic spectrum was derived from that for the observed spectrum by correcting for EBL attenuation.

The time-dependent, EBL-corrected energy flux values shown in Fig. 1 and reported in Extended Data Table 1 were computed with an analytical procedure. For each time bin, the value of the energy flux was computed as the integral between 0.3 and 1 TeV of the best-fit spectral power-law function derived with the forward-folding method. Accordingly, the errors were calculated analytically through standard procedures for error propagation, taking into account the covariance matrix. Moreover, the analytical results were checked against those computed with a toy Monte Carlo simulation, which gave comparable results.

The lower limits on the maximum event energy were computed by an iterative procedure in which a power-law model was assumed for the intrinsic spectrum and a different cut was applied to the maximum event energy for each iteration. For each value of the energy cut, a forward-folding fit was performed and a χ^2 value was obtained. The final result was obtained by finding the value of the energy cut for which the χ^2 variation corresponded to a given confidence level, set here to 95%.

The number of events in each time and energy bin shown in Fig. 3 was computed using the forward-folding EBL-corrected spectrum, the instrument effective area and the effective time of the observation. For the highest-energy bins, the corresponding numbers for the time interval between $T_0 + 62$ s and $T_0 + 1{,}227$ s are listed in Extended Data Table 2.

The number of observed excess events in bins of estimated energy are reported in Extended Data Table 3. Also listed are the expected number of photons in the same energy bins, obtained from the power-law model of the intrinsic spectrum by convolving it with the effect of EBL attenuation and the instrument response function for the zenith angles of this observation. We note that the counts in bins of estimated energy cannot be used to derive physical inferences. Spectral information that is physically meaningful must be computed as a function of the true energy of the events through an unfolding procedure using the energy migration matrix. Figure 2 shows such unfolded spectra (both intrinsic and observed) as a function of the true event energies.

Fermi-LAT data analysis for GRB 190114C

The publicly available Pass 8 (P8R3) LAT data for GRB 190114C were processed using the Conda Fermitools v1.0.2 package, distributed by the Fermi collaboration (https://fermi.gsfc.nasa.gov/ssc/data/analysis/software/). Events of the 'Transient' class (P8R3_TRANSIENT020_V2) were selected within 10° from the source position. We assumed a power-law spectrum in the 0.1–10 GeV energy range, also accounting for the diffuse Galactic and extragalactic backgrounds, as described in the analysis manual (https://fermi.gsfc.nasa.gov/ssc/data/analysis/scitools/). To compute the source fluxes, we first checked that the spectral index was consistent with −2 for the entire 62–180 s interval after T_0, and then repeated the fit, fixing the index to this value. The LAT energy flux shown in Fig. 1 was computed as the integral of the best-fit power-law model within the corresponding energy range.

XRT light curve

The XRT light curve shown in Fig. 1 was derived using the online analysis tool that is publicly available at the Swift-XRT repository (http://www.swift.ac.uk/xrt_curves/). The spectral data collected in the 'windowed timing' mode suffered from an instrumental effect, causing a non-physical excess of counts below ~0.8 keV (ref. [47]). To remove this effect, we considered the best-fit model of spectral data above 1 keV and estimated a conversion factor from the number of counts to deabsorbed flux equal to 10^{-10} erg cm^{-2} per count. To obtain the energy-flux light curve, we applied this conversion factor to the count rate as a function of time in the interval 62–2,000 s.

Synchrotron burnoff limit for the afterglow emission

GRB afterglows are triggered by external shocks that decelerate and dissipate their kinetic energy in the ambient medium, consequently

producing a nonthermal distribution of electrons via mechanisms such as shock acceleration[2]. The maximum energy of electrons that can be attained in the reference frame comoving with the post-shock region can be estimated by equating the timescales of acceleration, τ_{acc}, and energy loss, τ_{loss}; the latter is primarily due to synchrotron radiation[29]. These are expected to scale with the electron Lorentz factor, γ, and the magnetic field strength, B, as $\tau_{acc} \propto \gamma B^{-1}$ and $\tau_{loss} \propto \gamma^{-1}B^{-2}$, so that the maximum electron Lorentz factor is $\gamma_{max} \propto B^{-1/2}$. Thus, the maximum energy of synchrotron emission $\varepsilon_{syn,max} \propto B\gamma_{max}^2$ is independent of B. Its numerical value in the shock comoving frame is $\varepsilon'_{syn,max} \approx 50-100$ MeV, which is determined only from fundamental constants and a factor of order 1 that characterizes the uncertainties in the acceleration timescale. The observed spectrum of afterglow synchrotron emission is then expected to display a cutoff below the energy $\varepsilon_{syn,max} \approx 100\ \mathrm{MeV} \times [\Gamma_b(t)/(1+z)]$, which depends only on the time-dependent bulk Lorentz factor $\Gamma_b(t)$ of the external shock. To estimate $\varepsilon_{syn,max}$ and its evolution, we use the $\Gamma_b(t)$ values derived from solutions to the dynamical equations of the external shock[48]. The resulting curves for $\varepsilon_{syn,max}$ are shown for cases of a medium with constant density (n = constant) and a medium with a radial density profile of $n(R) = AR^{-2}$ (with $A = 3\times10^{35}A_*\ \mathrm{cm}^{-1}$, where A_* is a parameter characterizing the normalization of the density), expected when a dense stellar wind is produced by the progenitor star (dotted and dashed lines in Fig. 3, respectively). These curves have been derived assuming small values for the density ($n = 0.01$ and $A_* = 0.01$) and the efficiency of prompt emission ($\eta_\gamma = 1\%$), which imply a large value for the isotropic-equivalent blast-wave kinetic energy ($E_{k,aft} = E_{iso}(1-\eta_\gamma)/\eta_\gamma$), resulting in high values of $\varepsilon_{syn,max}$. Even with such extreme assumptions, the energy of photons detected by MAGIC are well above $\varepsilon_{syn,max}$ (Fig. 3).

Constraints on proton synchrotron afterglow emission

Synchrotron emission by protons accelerated to ultrahigh energies in the external shock has been proposed as a mechanism for gigaelectronvolt–teraelectronvolt emission in GRB afterglows, potentially at energies above the burnoff limit for electron synchrotron emission[10,12,13,49,50]. We discuss whether this process provides a viable explanation for the teraelectronvolt emission observed here, following the formulation of ref. [12]. For the case of a uniform external medium with density $n = n_0\ \mathrm{cm}^{-3}$, the maximum expected energy of proton synchrotron emission in the observer frame is

$$\varepsilon_{psyn,max} = (7.6\ \mathrm{GeV})\eta^{-2}\epsilon_B^{3/2}(n_0E_{k,53})^{3/4}t_s^{-1/4}(1+z)^{-3/4} \quad (1)$$

where $E_{k,aft} = 10^{53}E_{k,53}$ erg, t_s is the observer time after the burst in seconds, ϵ_B is the fraction of energy in magnetic fields relative to that dissipated behind the shock, and η is a factor of order 1 that characterizes the acceleration timescale. Even when assuming optimistic values of $\epsilon_B = 0.5$ and $\eta = 1$, realizing $\varepsilon_{psyn,max} \gtrsim 1$ TeV at $t \approx 100$ s for a GRB at $z = 0.42$ requires $n_0E_{k,53} \gtrsim 10^4$, which is a very high value for the product of the blastwave energy and the external medium density.

Even more severe is the requirement to reproduce the observed teraelectronvolt flux and spectrum. Assuming a power-law energy distribution with index $-p$ for the accelerated protons, their synchrotron emission is expected to have a single power-law spectrum with photon index $\alpha_{int} = -(p+1)/2$, extending from a minimum energy

$$\varepsilon_m = (3.7\times10^{-3}\ \mathrm{eV})\xi_p^{-2}\epsilon_p^2\epsilon_B^{1/2}E_{k,53}^{1/2}t_s^{-3/2}(1+z)^{1/2} \quad (2)$$

with differential energy flux

$$f(\varepsilon=\varepsilon_m) = (1.3\times10^{-28}\ \mathrm{erg\ cm^{-2}\ s^{-1}\ Hz^{-1}}) \times \xi_p\epsilon_B^{1/2}n_0^{1/2}E_{k,53}D_{28}^{-2}(1+z) \quad (3)$$

up to $\varepsilon = \varepsilon_{psyn,max}$, where ξ_p is the fraction of the number of protons swept up by the shock that are accelerated, ϵ_p is the fraction of the energy of the accelerated protons relative to that dissipated behind the shock, and $D = 10^{28}D_{28}$ cm is the luminosity distance of the GRB. The observed intrinsic spectral index $\alpha_{int} \approx -2$ at $t \approx 100$ s implies $p \approx 3$. If $p = 3$ and the spectrum extends to $\varepsilon = 1$ TeV without a cutoff, the energy flux at 1 TeV is

$$F(\varepsilon = 1\ \mathrm{TeV}) = (1.1\times10^{-16}\ \mathrm{erg\ cm^{-2}\ s^{-1}}) \times \epsilon_p^2\xi_p^{-1}\epsilon_B n_0^{1/2}E_{k,53}^{3/2}D_{28}^{-2}t_s^{-3/2}(1+z)^{3/2} \quad (4)$$

With optimistic assumptions of $\epsilon_B = 0.5$, $\eta = 1$, $\epsilon_p = 0.5$ and $\xi_p = 0.1$, accounting for the observed 0.3–1 TeV flux at $t \approx 100$ s of $F \approx 4\times10^{-8}\ \mathrm{erg\ cm^{-2}\ s^{-1}}$ necessitates $n_0^{1/2}E_{k,53}^{3/2} \gtrsim 10^{11}$. Even in the extreme case of a GRB occurring at the centre of a dense molecular cloud with $n = 10^6\ \mathrm{cm}^{-3}$, the blastwave energy must be $E_{k,aft} > 2\times10^{59}$ erg, greatly exceeding the energy available for any plausible GRB progenitor[2]. This conclusion is qualitatively valid regardless of how the electron synchrotron emission is modelled or whether the external medium has a density profile characteristic of a progenitor stellar wind. Although proton synchrotron emission may possibly explain the gigaelectronvolt emission observed in some GRBs[50], it is not favoured as the origin of the luminous teraelectronvolt emission observed in GRB 190114C, owing to its low radiative efficiency. A more plausible mechanism may be inverse Compton emission by accelerated electrons[9–11,51].

Past teraelectronvolt-band observations of GRBs with MAGIC and other facilities

Although the search for teraelectronvolt γ-rays from GRBs has continued over many years using a variety of experimental techniques, no clear detections had been previously achieved[52–63]. Designed with the primary goal of GRB follow-up observations, MAGIC has been responding to GRB alerts since 15 July 2004. For the first five years, MAGIC operated as a single telescope (MAGIC-I), reacting mainly to alerts from Swift. After the second telescope (MAGIC-II) was added in 2009, GRB observations have been carried out in stereoscopic mode. Excluding cases when useful data could not be taken owing to hardware problems or adverse weather conditions, 105 GRBs were observed from July 2004 to February 2019. Of these, 40 have determined redshifts, among which 8 and 3 have redshifts lower than 1 and 0.5, respectively. Observations started less than 30 min after the burst for 66 events (of which 33 lack redshifts) and less than 60 s for 14 events; the small number of events in the latter case is mainly due to bad weather conditions or observational criteria not being fulfilled at the time of the alert.

Despite 15 years of dedicated efforts, no unambiguous evidence for γ-ray signals from GRBs had been seen by MAGIC before GRB 190114C. The flux upper limits for GRBs observed in 2005–2006 were found to be consistent with simple power-law extrapolations of their low-energy spectra when EBL attenuation was taken into account[64]. More detailed studies were presented for GRB 080430[65] and GRB 090102[66], which were observed simultaneously with MAGIC and other instruments in different energy bands. Since 2013, GRB observations have been performed with the new automatic procedure described above[37,38]. In addition, for some bright GRBs detected by Fermi-LAT, late-time observations have been conducted up to one day after the burst to search for potential signals extended in time.

The case of GRB 190114C can be compared with other GRBs followed up by MAGIC under similar conditions. Aside from the intrinsic spectrum, the main factors affecting the detectability of a GRB by IACTs are the redshift z (stronger EBL attenuation for higher z), the zenith distance (higher energy threshold for higher zenith distance), the external light conditions and the delay time T_{delay} between the GRB and the beginning of the observations. If we select GRBs with $z < 1$ and $T_{delay} < 1$ h, only four events remain, as listed in Extended Data Table 5. Except for GRB 190114C, these are all short GRBs, which is not surprising as they are known to be distributed at redshifts appreciably lower than those of long GRBs[67]. A few other long GRBs with $z < 1$ and $T_{delay} < 1$ h were followed up by MAGIC, but the observations were not successful owing to technical problems or adverse observing conditions. There is also

a fair fraction of events without measured redshifts. Assuming that they follow the known z distribution of long GRBs, ~20% of the events are expected at $z<1$ (ref. [68]). Since 30 long GRBs without redshifts were observed by MAGIC with $T_{delay}<1$ h, only a few events with observing conditions and z similar to that of GRB 190114C are expected to be observed during the whole MAGIC GRB campaign .

A similar analysis for past GRBs observed by other Cherenkov telescopes is not possible, because not all of the relevant ancillary information is available. However, summaries of past efforts have been reported. Of the 150 GRBs followed up by VERITAS until February 2018[63], 50 had observations starting within 180 s from the satellite trigger time. H.E.S.S. also conducted several tens of GRB follow-up observations until 2017[58,69]. 64 GRBs were observed by HAWC[61] until February 2017. Milagrito and Milagro observed 54 GRBs from February 1997 to May 1998[70] and more than 130 GRBs from January 2000 to March 2008, respectively[71,72]. None of these considerable observational efforts provided any convincing detection, although some hints at low significance have been found. A case of particular interest was the Milagrito result for GRB 970417A[53], although its statistical significance was not high enough to fully rule out a background event.

Data availability

Raw data were generated at the MAGIC telescopes large-scale facility. Derived data supporting the findings of this study are available from the corresponding authors upon request. Source data for Figs. 1–3 are provided with the paper.

Code availability

Proprietary data reconstruction codes were generated at the MAGIC telescope large-scale facility. Information supporting the findings of this study is available from the corresponding authors upon request.

32. Amati, L. et al. Intrinsic spectra and energetics of BeppoSAX gamma-ray bursts with known redshifts. *Astron. Astrophys.* **390**, 81–89 (2002).
33. Yonetoku, D. et al. Gamma-ray burst formation rate inferred from the spectral peak energy-peak luminosity relation. *Astrophys. J.* **609**, 935–951 (2004).
34. Nava, L. et al. A complete sample of bright Swift long gamma-ray bursts: testing the spectral-energy correlations. *Mon. Not. R. Astron. Soc.* **421**, 1256–1264 (2012).
35. Vandenbroucke, J. GRB 190114C: Search for high-energy neutrinos with IceCube. *The Astronomer's Telegram* 12395 http://www.astronomerstelegram.org/?read=12395 (2019).
36. Barthelmy, S. GCN capabilities and status, and the incorporation of LIGO/Virgo. In *APS Meeting Abstracts*, M13.004 (APS, 2016).
37. Carosi, A. et al. Recent follow-up observations of GRBs in the very high energy band with the MAGIC Telescopes. In *34th International Cosmic Ray Conference (ICRC2015)* **34** (eds Borisov, A. S. et al.) 809 (POS, 2015).
38. MAGIC GRB group. Search for high energy emission from GRBs with MAGIC. *Proc. Int. Astron. Union* **12**, 70–73 (2016).
39. Fomin, V. et al. New methods of atmospheric Cherenkov imaging for gamma-ray astronomy. I. The false source method. *Astropart. Phys.* **2**, 137–150 (1994).
40. Ahnen, M. L. et al. Performance of the MAGIC telescopes under moonlight. *Astropart. Phys.* **94**, 29–41 (2017).
41. Fruck, C. et al. A novel LIDAR-based atmospheric calibration method for improving the data analysis of MAGIC. Preprint at https://arxiv.org/abs/1403.3591 (2014).
42. Li, T. P. & Ma, Y. Q. Analysis methods for results in gamma-ray astronomy. *Astrophys. J.* **272**, 317–324 (1983).
43. Schmelling, M. The method of reduced cross-entropy. A general approach to unfold probability distributions. *Nucl. Instrum. Methods Phys. Res. A* **340**, 400–412 (1994).
44. Franceschini, A., Rodighiero, G. & Vaccari, M. Extragalactic optical-infrared background radiation, its time evolution and the cosmic photon-photon opacity. *Astron. Astrophys.* **487**, 837–852 (2008).
45. Finke, J. D., Razzaque, S. & Dermer, C. D. Modeling the extragalactic background light from stars and dust. *Astrophys. J.* **712**, 238–249 (2010).
46. Gilmore, R. C., Somerville, R. S., Primack, J. R. & Domínguez, A. Semi-analytic modelling of the extragalactic background light and consequences for extragalactic gamma-ray spectra. *Mon. Not. R. Astron. Soc.* **422**, 3189–3207 (2012).
47. Beardmore, A. The Swift-XRT WT mode spectrum of GRB190114C. *GCN Circulars* 23736 https://gcn.gsfc.nasa.gov/gcn3/23736.gcn3 (2019).
48. Nava, L., Sironi, L., Ghisellini, G., Celotti, A. & Ghirlanda, G. Afterglow emission in gamma-ray bursts – I. Pair-enriched ambient medium and radiative blast waves. *Mon. Not. R. Astron. Soc.* **433**, 2107–2121 (2013).
49. Totani, T. Very strong TeV emission as gamma-ray burst afterglows. *Astrophys. J. Lett.* **502**, 13–16 (1998).
50. Razzaque, S. A leptonic–hadronic model for the afterglow of gamma-ray burst 090510. *Astrophys. J. Lett.* **724**, 109–112 (2010).
51. Galli, A. & Piro, L. Prospects for detection of very high-energy emission from GRB in the context of the external shock model. *Astron. Astrophys.* **489**, 1073–1077 (2008).
52. Connaughton, V. & VERITAS Collaboration. Gamma-ray bursts at VERITAS energies. *Astropart. Phys.* **11**, 255–257 (1999).
53. Atkins, R. et al. Evidence for TEV Emission from GRB 970417A. *Astrophys. J. Lett.* **533**, 119–122 (2000).
54. Atkins, R. et al. Limits on very high energy emission from gamma-ray bursts with the milagro observatory. *Astrophys. J. Lett.* **604**, 25–28 (2004).
55. Abdo, A. A. et al. Milagro constraints on very high energy emission from short-duration gamma-ray bursts. *Astrophys. J.* **666**, 361–367 (2007).
56. Horan, D. et al. Very high energy observations of gamma-ray burst locations with the Whipple telescope. *Astrophys. J.* **655**, 396–405 (2007).
57. Aharonian, F. et al. HESS observations of the prompt and afterglow phases of GRB 060602B. *Astrophys. J.* **690**, 1068–1073 (2009).
58. Aharonian, F. et al. HESS observations of γ-ray bursts in 2003–2007. *Astron. Astrophys.* **495**, 505–512 (2009).
59. Acciari, V. A. et al. VERITAS observations of gamma-ray bursts detected by Swift. *Astrophys. J.* **743**, 62 (2011).
60. H.E.S.S. Collaboration. Search for TeV gamma-ray emission from GRB 100621A, an extremely bright GRB in X-rays, with H.E.S.S. *Astron. Astrophys.* **565**, A16 (2014).
61. Alfaro, R. et al. Search for very-high-energy emission from gamma-ray bursts using the first 18 months of data from the HAWC gamma-ray observatory. *Astrophys. J.* **843**, 88 (2017).
62. Hoischen, C. et al. GRB Observations with H.E.S.S. II. In *International Cosmic Ray Conference (ICRC2017)* (2017); https://arxiv.org/abs/1708.01088.
63. Abeysekara, A. U. et al. A strong limit on the very-high-energy emission from GRB 150323A. *Astrophys. J.* **857**, 33 (2018).
64. Albert, J. et al. MAGIC upper limits on the very high energy emission from gamma-ray bursts. *Astrophys. J.* **667**, 358–366 (2007).
65. Aleksić, J. et al. MAGIC observation of the GRB 080430 afterglow. *Astron. Astrophys.* **517**, A5 (2010).
66. Aleksić, J. et al. MAGIC upper limits on the GRB 090102 afterglow. *Mon. Not. R. Astron. Soc.* **437**, 3103–3111 (2014).
67. Ghirlanda, G. et al. Short gamma-ray bursts at the dawn of the gravitational wave era. *Astron. Astrophys.* **594**, A84 (2016).
68. Perley, D. A. et al. The Swift gamma-ray burst host galaxy legacy survey. I. Sample selection and redshift distribution. *Astrophys. J.* **817**, 7 (2016).
69. Lennarz, D. et al. Searching for TeV emission from GRBs: the status of the H.E.S.S. GRB programme. Preprint at https://arxiv.org/abs/1307.6897 (2013).
70. Smith, A. J. et al. Results from the Milagrito experiment. *AIP Conf. Proc.* **515**, 441–447 (2000).
71. Aune, T. A search for GeV–TeV emission from GRBs using the Milagro detector. *AIP Conf. Proc.* **1133**, 385–387 (2009).
72. Saz Parkinson, P. M. A search for GeV–TeV emission from gamma-ray bursts using the Milagro detector. *AIP Conf. Proc.* **1112**, 181–186 (2009).
73. Evans, P. A. et al. The Swift Burst Analyser. I. BAT and XRT spectral and flux evolution of gamma-ray bursts. *Astron. Astrophys.* **519**, A102 (2010).

Acknowledgements We are grateful to G. Sinnis for remarks that helped us improve the format and content of this manuscript. We dedicate this paper to the memory of E. Lorenz. With his innovative spirit, infinite enthusiasm and vast knowledge of experimental methods, techniques and materials, he played a key role in optimizing the design of MAGIC, specifically for observations of GRBs. We thank the Instituto de Astrofísica de Canarias for the excellent working conditions at the Observatorio del Roque de los Muchachos in La Palma. We acknowledge financial support by the German BMBF and MPG, the Italian INFN and INAF, the Swiss National Fund (SNF), the ERDF under the Spanish Ministry of Economy and Competitiveness (FPA2017-87859-P, FPA2017-85668-P,FPA2017-82729-C6-2-R, FPA2017-82729-C6-6-R, FPA2017-82729-C6-5-R, AYA2015-71042-P,AYA2016-76012-C3-1-P,ESP2017-87055-C2-2-P,FPA2017-90566-REDC), the Indian Department of Atomic Energy, the Japanese JSPS and MEXT, the Bulgarian Ministry of Education and Science, National RI Roadmap Project DO1-153/28.08.2018, and the Academy of Finland for grant number 320045. This work was also supported by the Spanish Centro de Excelencia 'Severo Ochoa' SEV-2016-0588 and SEV-2015-0548 and Unidad de Excelencia 'María de Maeztu' MDM-2014-0369, by the Croatian Science Foundation (HrZZ) Project IP-2016-06-9782 and the University of Rijeka Project 13.12.1.3.02, by the DFG Collaborative Research Centers SFB823/C4 and SFB876/C3, the Polish National Research Centre grant UMO-2016/22/M/ST9/00382, and by the Brazilian MCTIC, CNPq and FAPERJ. S.I. is supported by JSPS KAKENHI grant number JP17K05460, MEXT, Japan, the RIKEN iTHEMS programme and the joint research programme of ICRR, University of Tokyo. L. Nava acknowledges funding from the European Union's Horizon 2020 Research and Innovation programme under the Marie Skłodowska-Curie grant agreement number 664931. K. Noda is supported by JSPS KAKENHI grant number JP19K21043, MEXT, Japan. A. Berti acknowledges support from the Physics Department of the University of Torino (through funding from the Department of Excellence) and from the Torino division of the Italian INFN. E. Moretti acknowledges funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement number 665919.

Author contributions The MAGIC telescope system was designed and constructed by the MAGIC Collaboration. The operation, data processing, calibration, Monte Carlo simulations of the detector and of theoretical models, and data analyses were performed by the members of the MAGIC Collaboration, who also discussed and approved the scientific results. All MAGIC collaborators contributed to the editing and comments to the final version of the manuscript.

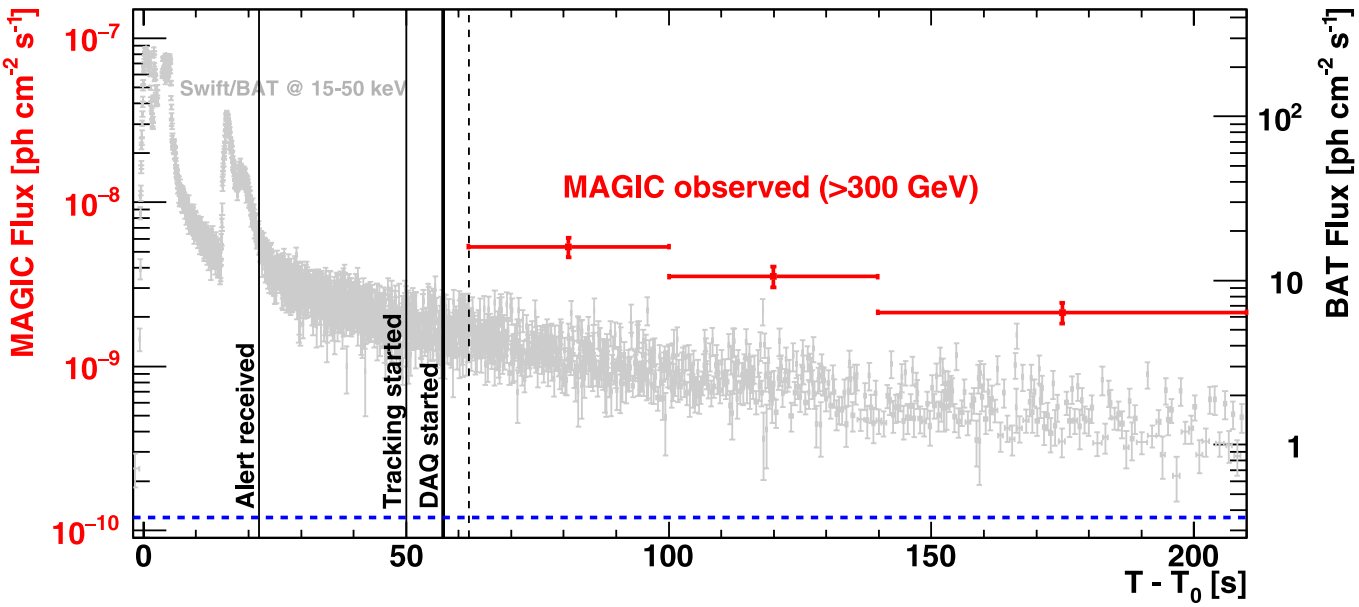
S.I. and L. Nava coordinated the interpretation of the data and, together with S. Covino, wrote the corresponding sections and contributed to the structuring and editing of the rest of the paper. K. Noda and A. Berti coordinated the analysis of the MAGIC data; together with E. Moretti they contributed to the analysis and the writing of the relevant sections. I.V. performed the Fermi-LAT analysis and, together with D. Miceli contributed to the calculation of limits, excesses and the curves in Fig. 3. R.M. contributed to coordinating, structuring and editing this paper.

Competing interests The authors declare no competing interests.

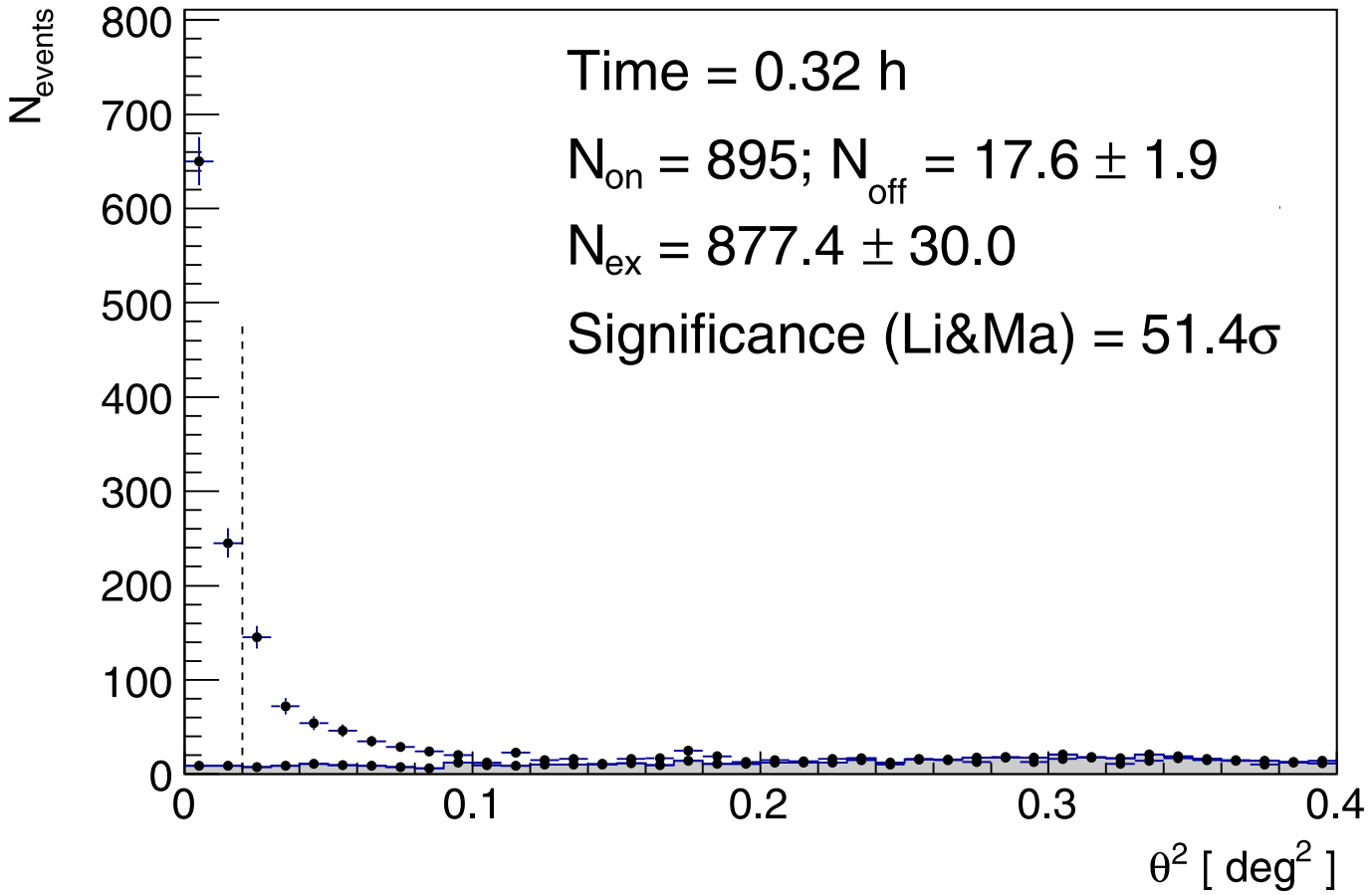
Additional information

Correspondence and requests for materials should be addressed to V.A.A.

Peer review information *Nature* thanks Gus Sinnis and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



Extended Data Fig. 1 | Light curves in the teraelectronvolt and kiloelectronvolt bands for GRB 190114C. Photon flux light curve above 0.3 TeV measured by MAGIC (red; from T_0 + 62 s to T_0 + 210 s), compared with that between 15 keV and 50 keV measured by Swift-BAT[73] (grey; from T_0 to T_0 + 210 s) and the photon flux above 0.3 TeV of the Crab Nebula (blue dashed line). The errors on the MAGIC photon fluxes correspond to one standard deviation. Vertical lines indicate the times when the alert was received (T_0 + 22 s) by MAGIC, when the tracking of the GRB by the telescopes started (T_0 + 50 s), when the data acquisition started (T_0 + 57 s), and when the data acquisition system (DAQ) became stable (T_0 + 62 s; dotted line).



Extended Data Fig. 2 | Significance of the γ-ray signal between T_0 + 62 s and T_0 + 1,227 s for GRB 190114C. Distribution of the squared angular distance, θ^2, for the MAGIC data (points) and background events (grey shaded area). θ^2 is defined as the squared angular distance between the nominal position of the source and the reconstructed arrival direction of the events. The dashed vertical line represents the value of the cut on θ^2. This defines the signal region, where the number of events coming from the source (N_{on}) and from the background (N_{off}) are computed. The errors for 'on' events are derived from Poissonian statistics. From N_{on} and N_{off}, the number of excess events (N_{ex}) is computed. The significance is calculated using the Li & Ma method[42].

Extended Data Table 1 | Energy flux between 0.3 and 1 TeV in selected time bins for GRB 190114C

Time bin [seconds after T_0]	Energy flux [erg cm^{-2} s^{-1}]	Spectral index
$62 - 100$	$[\,5.64 \pm 0.90\,(\text{stat})\,^{+3.24}_{-3.22}\,(\text{sys})\,] \cdot 10^{-8}$	$-1.86\,^{+0.36}_{-0.40}\,(\text{stat})\,^{+0.12}_{-0.21}\,(\text{sys})$
$100 - 140$	$[\,3.31 \pm 0.67\,(\text{stat})\,^{+2.71}_{-1.84}\,(\text{sys})\,] \cdot 10^{-8}$	$-2.15\,^{+0.43}_{-0.48}\,(\text{stat})\,^{+0.25}_{-0.32}\,(\text{sys})$
$140 - 210$	$[\,1.89 \pm 0.36\,(\text{stat})\,^{+1.72}_{-0.94}\,(\text{sys})\,] \cdot 10^{-8}$	$-2.31\,^{+0.47}_{-0.54}\,(\text{stat})\,^{+0.15}_{-0.22}\,(\text{sys})$
$210 - 361.5$	$[\,7.54 \pm 1.60\,(\text{stat})\,^{+6.46}_{-4.41}\,(\text{sys})\,] \cdot 10^{-9}$	$-2.53\,^{+0.53}_{-0.62}\,(\text{stat})\,^{+0.22}_{-0.24}\,(\text{sys})$
$361.5 - 800$	$[\,3.10 \pm 0.70\,(\text{stat})\,^{+1.20}_{-2.36}\,(\text{sys})\,] \cdot 10^{-9}$	$-2.41\,^{+0.51}_{-0.65}\,(\text{stat})\,^{+0.27}_{-0.34}\,(\text{sys})$
$800 - 2454$	$[\,4.54 \pm 2.04\,(\text{stat})\,^{+7.66}_{-1.96}\,(\text{sys})\,] \cdot 10^{-10}$	$-3.10\,^{+0.87}_{-1.25}\,(\text{stat})\,^{+0.75}_{-0.24}\,(\text{sys})$
$62 - 2454$ (time integrated)	-	$-2.22\,^{+0.23}_{-0.25}\,(\text{stat})\,^{+0.21}_{-0.26}\,(\text{sys})$

Values listed correspond to the light curve in Fig. 1. For each time bin, columns represent the start and end time of the bin, the EBL-corrected energy flux in the 0.3–1 TeV range, and the best-fit spectral photon indices. The last row reports the value of the intrinsic spectral index for the time-integrated spectrum (Fig. 2). The reported statistical errors (stat) correspond to one standard deviation, whereas systematic errors (sys) are derived from the variation of the light scale by ±15% (see Methods).

Extended Data Table 2 | Number of γ-rays from GRB 190114C in the highest-energy bins

E_{min} [TeV]	E_{max} [TeV]	Model counts in [E_{min}; E_{max}]	Significance above E_{min}
0.71	1.10	25.4	5.8
1.10	1.70	4.1	2.5
1.70	2.64	0.9	1.5
2.64	4.09	0.1	0.1

The number of γ-ray counts was estimated from the MAGIC data using the power-law spectral model for the time interval between T_0 + 62 s and T_0 + 1,227 s.

Extended Data Table 3 | Observed and expected number of events in estimated-energy bins for GRB 190114C

$E_{\mathrm{est,min}}$ [TeV]	$E_{\mathrm{est,max}}$ [TeV]	Observed photons	Expected photons
0.19	0.29	155 ± 13	219 ± 73
0.29	0.46	598 ± 26	564 ± 53
0.46	0.71	154 ± 13	180 ± 16
0.71	1.10	32 ± 6	28 ± 3
1.10	1.70	6.0 ± 2.9	5.6 ± 0.4
1.70	2.64	2.3 ± 1.8	1.2 ± 0.1

The number of expected events is calculated from the intrinsic spectrum power-law model, by convolving it with the effect of EBL attenuation and the instrument response function of the telescope for these large zenith angles. The energy binning in estimated energy matches the one in true energy (after unfolding) shown in Fig. 2 and Extended Data Table 2. The large uncertainty in the number of expected events in the lowest-energy bin is dominated by the uncertainty in the very low effective area of the telescopes close to the energy threshold of this analysis. The numbers reported in this table cannot be used directly for any physical inference. The measured spectrum needs to be first unfolded using the energy migration matrix[21].

Extended Data Table 4 | Spectral indices for different EBL models

Time bin [seconds after T_0]	D11	F08	FI10	G12
62 – 100	$-1.86^{+0.36}_{-0.40}$	$-2.04^{+0.36}_{-0.40}$	$-1.81^{+0.36}_{-0.40}$	$-1.95^{+0.36}_{-0.39}$
100 – 140	$-2.15^{+0.43}_{-0.48}$	$-2.32^{+0.43}_{-0.48}$	$-2.09^{+0.43}_{-0.48}$	$-2.23^{+0.42}_{-0.48}$
140 – 210	$-2.31^{+0.47}_{-0.54}$	$-2.48^{+0.47}_{-0.54}$	$-2.25^{+0.47}_{-0.54}$	$-2.39^{+0.47}_{-0.53}$
210 – 361.5	$-2.53^{+0.53}_{-0.62}$	$-2.69^{+0.52}_{-0.61}$	$-2.46^{+0.52}_{-0.61}$	$-2.60^{+0.52}_{-0.61}$
361.5 – 800	$-2.41^{+0.51}_{-0.65}$	$-2.58^{+0.51}_{-0.64}$	$-2.34^{+0.51}_{-0.64}$	$-2.49^{+0.51}_{-0.64}$
800 – 2454	$-3.10^{+0.87}_{-1.25}$	$-3.20^{+0.83}_{-1.20}$	$-2.96^{+0.83}_{-1.20}$	$-3.08^{+0.82}_{-1.19}$
62 – 2454 (time integrated)	$-2.22^{+0.23}_{-0.25}$	$-2.39^{+0.23}_{-0.25}$	$-2.15^{+0.23}_{-0.25}$	$-2.29^{+0.23}_{-0.24}$

The abbreviations refer to the different EBL model adopted in each case. D11: Dominguez et al.[25] (reported also in Extended Data Table 1); F08: Franceschini et al.[44]; FI10: Finke et al.[45]; G12: Gilmore et al.[46]. The errors correspond to one standard deviation.

Extended Data Table 5 | List of GRBs observed under adequate technical and weather conditions by MAGIC with $z < 1$ and $T_{delay} < 1$ h

Event	redshift	T_{delay} (s)	Zenith angle (deg)
GRB 061217	0.83	786.0	59.9
GRB 100816A	0.80	1439.0	26.0
GRB 160821B	0.16	24.0	34.0
GRB 190114C	0.42	58.0	55.8

The zenith angle at the beginning of the observations is reported in the last column. All GRBs except GRB 061217 were observed in stereoscopic mode. GRB 061217, GRB 100816A and GRB 160821B are short GRBs, whereas GRB 190114C is a long GRB. Observations of a few other long GRBs with the same criteria were also conducted but are not listed here, because they were affected by technical problems or adverse observing conditions.

Observation of inverse Compton emission from a long γ-ray burst

https://doi.org/10.1038/s41586-019-1754-6

Received: 20 July 2019

Accepted: 18 October 2019

Published online: 20 November 2019

A list of authors and affiliations appears at the end of the paper.

Long-duration γ-ray bursts (GRBs) originate from ultra-relativistic jets launched from the collapsing cores of dying massive stars. They are characterized by an initial phase of bright and highly variable radiation in the kiloelectronvolt-to-megaelectronvolt band, which is probably produced within the jet and lasts from milliseconds to minutes, known as the prompt emission[1,2]. Subsequently, the interaction of the jet with the surrounding medium generates shock waves that are responsible for the afterglow emission, which lasts from days to months and occurs over a broad energy range from the radio to the gigaelectronvolt bands[1–6]. The afterglow emission is generally well explained as synchrotron radiation emitted by electrons accelerated by the external shock[7–9]. Recently, intense long-lasting emission between 0.2 and 1 teraelectronvolts was observed from GRB 190114C[10,11]. Here we report multi-frequency observations of GRB 190114C, and study the evolution in time of the GRB emission across 17 orders of magnitude in energy, from 5×10^{-6} to 10^{12} electronvolts. We find that the broadband spectral energy distribution is double-peaked, with the teraelectronvolt emission constituting a distinct spectral component with power comparable to the synchrotron component. This component is associated with the afterglow and is satisfactorily explained by inverse Compton up-scattering of synchrotron photons by high-energy electrons. We find that the conditions required to account for the observed teraelectronvolt component are typical for GRBs, supporting the possibility that inverse Compton emission is commonly produced in GRBs.

On 14 January 2019, following an alert from the Neil Gehrels Swift Observatory (hereafter Swift) and the Fermi satellite, the Major Atmospheric Gamma Imaging Cherenkov (MAGIC) telescopes observed and detected radiation up to at least 1 TeV from GRB 190114C. Before the MAGIC detection, GRB emission had only been reported at much lower energies, below 100 GeV, first by CGRO-EGRET and more recently by AGILE-GRID and Fermi-LAT (see ref. [12] for a recent review).

Detection of teraelectronvolt radiation opens a new window in the electromagnetic spectrum for the study of GRBs[10]. Its announcement[13] triggered an extensive campaign of follow-up observations. Owing to the relatively low redshift of $z = 0.4245 \pm 0.0005$ (Methods) of the GRB (corresponding to a luminosity distance of about 2.3 Gpc), a comprehensive set of multi-wavelength data could be collected. We present observations gathered from instruments onboard six satellites and 15 ground telescopes (radio, submillimetre, near-infrared (NIR), optical, ultraviolet (UV), and very-high-energy γ-rays; see Methods) for the first ten days after the burst. The frequency range covered by these observations spans more than 17 orders of magnitude, from 1 to about 2×10^{17} GHz, the most extensive so far for a GRB. The light curves of GRB 190114C at different frequencies are shown in Fig. 1.

The prompt emission of GRB 190114C was simultaneously observed by several space missions covering the spectral range from 8 keV to about 100 GeV (Methods). The prompt light curve shows a complex temporal structure with several emission peaks (Methods, Extended Data Fig. 1), with a total duration of about 25 s (see dashed line in Fig. 1) and total radiated energy of $E_{\gamma,\mathrm{iso}} = (2.5 \pm 0.1) \times 10^{53}$ erg (isotropic equivalent; 1 erg = 10^{-7} J) in the energy range 1–10^4 keV (ref. [14]). During the time of inter-burst quiescence, at $t \approx 5$–15 s, and after the end of the last prompt pulse, at $t \gtrsim 25$ s, the flux decays smoothly, following a power law of $F \propto t^{\alpha}$ as a function of time t with $\alpha_{10-1{,}000\mathrm{keV}} = -1.10 \pm 0.01$ (ref. [14]). The temporal and spectral characteristics of this smoothly varying component support an interpretation in terms of afterglow synchrotron radiation, making this one of the few clear cases of afterglow emission detected in the band 10–10^4 keV during the prompt-emission phase. The onset of the afterglow component is then estimated to occur around $t \approx 5$–10 s (refs. [14,15]), implying an initial bulk Lorentz factor between 300 and 700 (Methods).

After about one minute from the start of the prompt emission, two additional high-energy telescopes began observations: MAGIC and Swift-XRT. The XRT (1–10 keV; blue data points in Fig. 1) and MAGIC (0.3–1 TeV; green data points in Fig. 1) light curves decay with time as a power law with decay indices of $\alpha_{\mathrm{X}} \approx -1.36 \pm 0.02$ and $\alpha_{\mathrm{TeV}} \approx -1.51 \pm 0.04$, respectively. The 0.3–1-TeV light curve shown in Fig. 1 was obtained after correcting for attenuation by the extragalactic background light (EBL)[10]. The teraelectronvolt-band emission is observable until about 40 min—much longer than the nominal duration of the prompt-emission phase. The NIR–optical light curves (square symbols) show a more complex behaviour. Initially, a fast decay is seen, where the emission is probably dominated by the reverse-shock component[16]. This is followed by a shallower decay, and subsequently a faster decay at $t \gtrsim 10^5$ s. The latter may indicate that the characteristic synchrotron frequency ν_{m} crosses the optical band (Extended Data Fig. 6), which is not atypical,

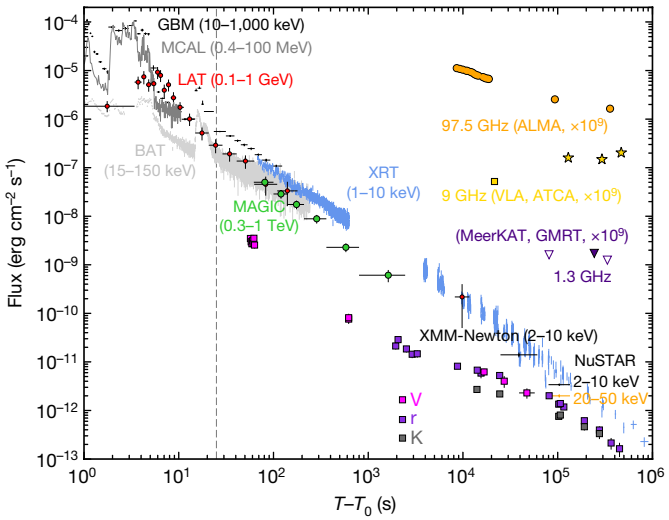


Fig. 1 | Multi-wavelength light curves of GRB 190114C. Energy flux at different wavelengths, from radio to γ-rays, versus time after the BAT trigger, at T_0 = 20:57:03.19 universal time (UT) on 14 January 2019. The light curve for the energy range 0.3–1 TeV (green circles) is compared with light curves at lower frequencies. Those for VLA (yellow square), ATCA (yellow stars), ALMA (orange circles), GMRT (purple filled triangle) and MeerKAT (purple open triangles) have been multiplied by 10^9 for clarity. The vertical dashed line marks approximately the end of the prompt-emission phase, identified as the end of the last flaring episode. For the data points, vertical bars show the 1σ errors on the flux, and horizontal bars represent the duration of the observation. The fluxes in the V, r and K filters (pink, purple and grey filled squares, respectively) have been corrected for extinction in the host and in our Galaxy; the contribution from the host galaxy has been subtracted.

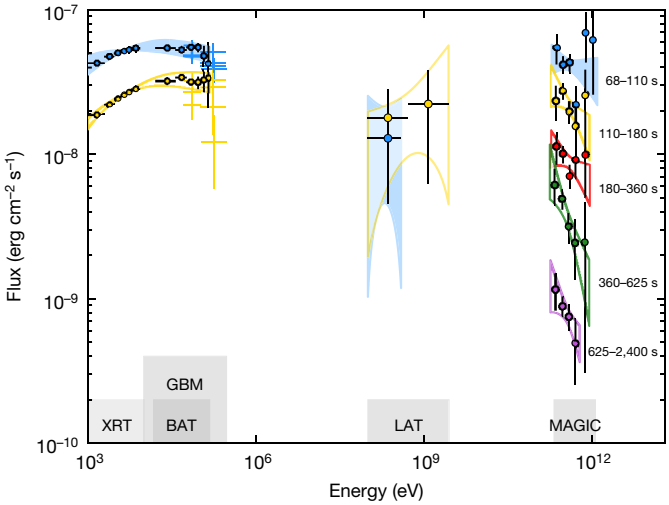


Fig. 2 | Multi-band spectra in the time interval 68–2,400 s. Five time intervals are considered: 68–110 s (blue), 110–180 s (yellow), 180–360 s (red), 360–625 s (green) and 625–2,400 s (purple). MAGIC data points have been corrected for attenuation caused by the EBL. Data from other instruments (Swift-XRT, Swift-BAT, Fermi-GBM and Fermi-LAT) are shown for the first two time intervals. For each time interval, LAT contour regions are shown, limiting the energy to the range in which photons are detected. MAGIC and LAT contour regions are drawn from the 1σ error of their best-fit power-law functions. For Swift data, the regions show the 90% confidence contours for the joint fit for XRT and BAT, obtained by fitting a smoothly broken power law to the data. Filled regions are used for the first time interval (68–110 s).

but usually occurs at earlier times. The relatively late time at which the break appears in GRB 190114C would then imply a very large value of ν_m, placing it in the X-ray band at about 10^2 s. The millimetre light curves (orange symbols) also show an initial fast decay in which the emission is dominated by the reverse shock, followed by emission at late times with nearly constant flux (Extended Data Fig. 3).

The spectral energy distributions (SEDs) of the radiation detected by MAGIC are shown in Fig. 2, where the whole duration of the emission detected by MAGIC is divided into five time intervals. For the first two time intervals, observations in the gigaelectronvolt and X-ray bands are also available. During the first time interval (68–110 s; blue data points and blue confidence regions), Swift-XRT, Swift-BAT and Fermi-GBM data show that the afterglow synchrotron component peaks in the X-ray band. At higher energies, up to 1 GeV, the SED is a decreasing function of energy, as supported by the Fermi-LAT flux between 0.1 and 0.4 GeV (Methods). On the other hand, at even higher energies, the MAGIC flux above 0.2 TeV implies a spectral hardening. This evidence is independent of the EBL model adopted to correct for the attenuation (Methods). This demonstrates that the newly discovered teraelectronvolt radiation is not a simple extension of the known afterglow synchrotron emission, but a separate spectral component.

The extended duration and the smooth, power-law temporal decay of the radiation detected by MAGIC (see green data points in Fig. 1) suggest an intimate connection between the teraelectronvolt emission and the broadband afterglow emission. The most natural candidate is synchrotron self-Compton (SSC) radiation in the external forward shock: the same population of relativistic electrons responsible for the afterglow synchrotron emission Compton up-scatters the synchrotron photons, leading to a second spectral component that peaks at higher energies. Teraelectronvolt afterglow emission can also be produced by hadronic processes, such as synchrotron radiation by protons accelerated to ultrahigh energies in the forward shock[17–19]. However, owing to their typically low radiation efficiency[6], reproducing the luminous teraelectronvolt emission observed here by such processes would imply unrealistically large power of accelerated protons[10]. Teraelectronvolt photons can also be produced via the SSC mechanism in internal shock synchrotron models of the prompt emission. However, numerical modelling (Methods) shows that prompt SSC radiation can account at most for a limited fraction (≲20%) of the observed teraelectronvolt flux, and only at early times ($t \lesssim 100$ s). Henceforth, we focus on the SSC process in the afterglow.

SSC emission has been predicted for GRB afterglows[9,12,18,20–27]. However, its quantitative significance has been uncertain because the SSC luminosity and spectral properties depend strongly on the poorly constrained physical conditions in the emission region (for example, the magnetic field strength). The detection of the teraelectronvolt component in GRB 190114C and the availability of multi-band observations offer the opportunity to investigate the relevant physics at a deeper level. SSC radiation may have been already detected in very bright GRBs, such as GRB 130427A, in which photons with energies of 10–100 GeV are challenging to explain by synchrotron processes, suggesting a different origin[28–30].

We model the full dataset (from the radio band to teraelectronvolt energies, for the first week after the explosion) as synchrotron plus SSC radiation, within the framework of the theory of afterglow emission from external forward shocks. The detailed modelling of the broadband emission and its evolution with time is presented in Methods. We discuss here the implications for the emission at $t < 2{,}400$ s and energies above >1 keV.

The soft spectra in the 0.2–1-TeV energy range (photon index $\Gamma_{\rm TeV} < -2$; see Extended Data Table 1) constrain the peak of the SSC component to below this energy range. The relatively small ratio between the spectral peak energies of the SSC ($E_{\rm p}^{\rm SSC} \lesssim 200$ GeV) and synchrotron ($E_{\rm p}^{\rm syn} \approx 10$ keV) components implies a relatively low value for the electron Lorentz factor ($\gamma \approx 2 \times 10^3$). This value is hard to reconcile with the

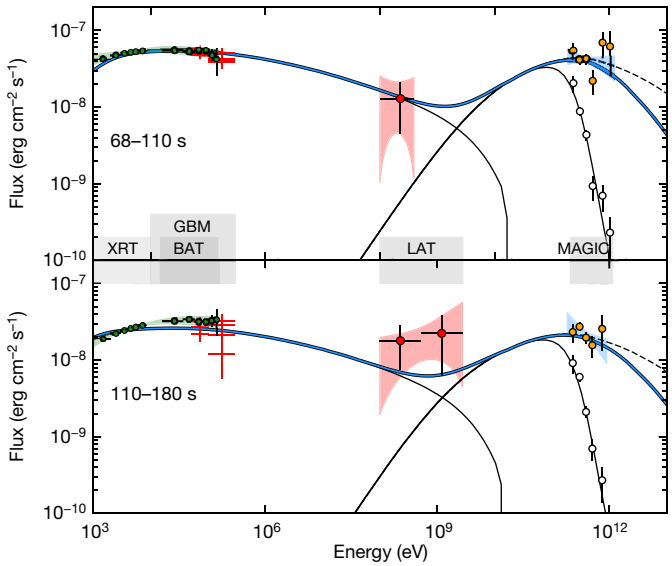


Fig. 3 | Modelling of the broadband spectra in the time intervals 68–110 s and 110–180 s. Thick blue curve, modelling of the multi-band data in the synchrotron and SSC afterglow scenario. Thin solid lines, synchrotron and SSC (observed spectrum) components. Dashed lines, SSC when internal γ–γ opacity is neglected. The adopted parameters are: $s = 0$, $\varepsilon_e = 0.07$, $\varepsilon_B = 8 \times 10^{-5}$, $p = 2.6$, $n_0 = 0.5$ and $E_k = 8 \times 10^{53}$ erg; see Methods. Empty circles show the observed MAGIC spectrum, that is, uncorrected for attenuation caused by the EBL. Contour regions and data points are as in Fig. 2.

observation of the synchrotron peak at energies higher than kiloelectronvolt. To explain the soft spectrum detected by MAGIC, it is necessary to invoke scattering in the Klein–Nishina regime for the electrons radiating at the spectral peak, as well as internal γ–γ absorption[31]. Although both of these effects tend to become less important with time, the spectral index in the 0.2–1-TeV band remains constant in time (or possibly evolves to softer values; Extended Data Table 1). This implies that the SSC peak energy moves to lower energies and crosses the MAGIC energy band. The energy at which attenuation by internal pair production becomes important indicates that the bulk Lorentz factor is about 140–160 at 100 s.

An example of the theoretical modelling in this scenario is shown in Fig. 3 (blue solid curve; see Methods for details). The dashed line shows the SSC spectrum when internal absorption is neglected. The thin solid line shows the model spectrum including EBL attenuation, in comparison to the MAGIC observations (empty circles).

We find that acceptable models of the broadband SED can be obtained if the conditions at the source are the following. The initial kinetic energy of the blast wave is $E_k \gtrsim 3 \times 10^{53}$ erg (isotropic-equivalent). The electrons swept up from the external medium are efficiently injected into the acceleration process and carry a fraction $\varepsilon_e \approx 0.05$–$0.15$ of the energy dissipated at the shock. The acceleration mechanism produces an electron population characterized by a non-thermal energy distribution, described by a power law with index $p \approx 2.4$–2.6, an injection Lorentz factor of $\gamma_m = (0.8\text{–}2) \times 10^4$ and a maximum Lorentz factor of $\gamma_{max} \approx 10^8$ (at about 100 s). The magnetic field behind the shock conveys a fraction $\varepsilon_B \approx (0.05\text{–}1) \times 10^{-3}$ of the dissipated energy. At $t \approx 100$ s, corresponding to a distance from the central engine of $R \approx (8\text{–}20) \times 10^{16}$ cm, the density of the external medium is $n \approx 0.5$–$5\ \mathrm{cm}^{-3}$ and the magnetic field strength is $B \approx 0.5$–5 G. The latter implies that the magnetic field was efficiently amplified from values of a few microgauss, which are typical of the unshocked ambient medium, owing to plasma instabilities or other mechanisms[6]. Not surprisingly, we find that $\varepsilon_e \gg \varepsilon_B$, which is a necessary condition for the efficient production of SSC radiation[18,20].

The blast-wave energy inferred from the modelling is comparable to the amount of energy released in the form of radiation during the prompt phase. The prompt-emission mechanism must then have dissipated and radiated no more than half of the initial jet energy, leaving the rest for the afterglow phase. The modelling of the multi-band data also allows us to infer how the total energy is shared between the synchrotron and SSC components. The resultant powers of the two components are comparable. We estimate that the energy in the synchrotron and SSC component are about 1.5×10^{52} erg and around 6.0×10^{51} erg, respectively, in the time interval 68–110 s, and about 1.3×10^{52} erg and around 5.4×10^{51} erg, respectively, in the time interval 110–180 s. Thus, previous studies of GRBs may have been missing a substantial fraction of the energy emitted during the afterglow phase that is essential to its understanding.

Finally, we note that the values of the afterglow parameters inferred from the modelling fall within the range of values typically inferred from broadband (radio to gigaelectronvolt) studies of GRB afterglow emission. This points to the possibility that SSC emission in GRBs may be a relatively common process that does not require special conditions to be produced, and its power is similar to that of synchrotron radiation.

The SSC component may then be detectable at teraelectronvolt energies in other relatively energetic GRBs, as long as the redshift is low enough to avoid severe attenuation by the EBL. This also provides support to earlier indications for SSC emission at gigaelectronvolt energies[28–30].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1754-6.

1. Mészáros, P. Theories of gamma-ray bursts. *Annu. Rev. Astron. Astrophys.* **40**, 137–169 (2002).
2. Piran, T. The physics of gamma-ray bursts. *Rev. Mod. Phys.* **76**, 1143–1210 (2005).
3. van Paradijs, J., Kouveliotou, C. & Wijers, R. A. M. J. Gamma-ray burst afterglows. *Annu. Rev. Astron. Astrophys.* **38**, 379–425 (2000).
4. Gehrels, N., Ramirez-Ruiz, E. & Fox, D. B. Gamma-ray bursts in the Swift era. *Annu. Rev. Astron. Astrophys.* **47**, 567–617 (2009).
5. Gehrels, N. & Mészáros, P. Gamma-ray bursts. *Science* **337**, 932–936 (2012).
6. Kumar, P. & Zhang, B. The physics of gamma-ray bursts & relativistic jets. *Phys. Rep.* **561**, 1–109 (2015).
7. Sari, R., Piran, T. & Narayan, R. Spectra and light curves of gamma-ray burst afterglows. *Astrophys. J. Lett.* **497**, 17–20 (1998).
8. Granot, J. & Sari, R. The shape of spectral breaks in gamma-ray burst afterglows. *Astrophys. J.* **568**, 820–829 (2002).
9. Mészáros, P. & Rees, M. J. Delayed GeV emission from cosmological gamma-ray bursts – impact of a relativistic wind on external matter. *Mon. Not. R. Astron. Soc.* **269**, L41–L43 (1994).
10. MAGIC Collaboration. Teraelectronvolt emission from the γ-ray burst GRB 190114C. *Nature* https://doi.org/10.1038/s41586-019-1750-x (2019).
11. Mirzoyan, R. et al. MAGIC detects the GRB 190114C in the TeV energy domain. *GCN Circulars* 23701 https//gcn.gsfc.nasa.gov/gcn3/23701.gcn3 (2019).
12. Nava, L. High-energy emission from gamma-ray bursts. *Int. J. Mod. Phys. D* **27**, 1842003 (2018).
13. Mirzoyan, R. et al. MAGIC detects the GRB 190114C. *The Astronomer's Telegram* 12390 http://www.astronomerstelegram.org/?read=12390 (2019).
14. Ajello, M. et al. Fermi and Swift observations of GRB 190114C: tracing the evolution of high-energy emission from prompt to afterglow. Preprint at https://arxiv.org/abs/1909.10605 (2019).
15. Ravasio, M. E. et al. GRB 190114C: from prompt to afterglow? *Astron. Astrophys.* **626**, A12 (2019).
16. Laskar, T. et al. ALMA detection of a linearly polarized reverse shock in GRB 190114C. *Astrophys. J. Lett.* **878**, 26 (2019).
17. Vietri, M. GeV photons from ultrahigh energy cosmic rays accelerated in gamma ray bursts. *Phys. Rev. Lett.* **78**, 4328–4331 (1997).
18. Zhang, B. & Mészáros, P. High-energy spectral components in gamma-ray burst afterglows. *Astrophys. J.* **559**, 110–122 (2001).
19. Razzaque, S. A leptonic–hadronic model for the afterglow of gamma-ray burst 090510. *Astrophys. J. Lett.* **724**, 109–112 (2010).
20. Sari, R. & Esin, A. A. On the synchrotron self-Compton emission from relativistic shocks and its implications for gamma-ray burst afterglows. *Astrophys. J.* **548**, 787–799 (2001).

21. Mészáros, P., Razzaque, S. & Zhang, B. GeV–TeV emission from γ-ray bursts. *New Astron. Rev.* **48**, 445–451 (2004).
22. Lemoine, M. The synchrotron self-Compton spectrum of relativistic blast waves at large *Y*. *Mon. Not. R. Astron. Soc.* **453**, 3772–3784 (2015).
23. Fan, Y.-Z. & Piran, T. High-energy γ-ray emission from gamma-ray bursts – before GLAST. *Front. Phys. China* **3**, 306–330 (2008).
24. Galli, A. & Piro, L. Prospects for detection of very high-energy emission from GRB in the context of the external shock model. *Astron. Astrophys.* **489**, 1073–1077 (2008).
25. Nakar, E., Ando, S. & Sari, R. Klein–Nishina effects on optically thin synchrotron and synchrotron self-Compton spectrum. *Astrophys. J.* **703**, 675–691 (2009).
26. Xue, R. R. et al. Very high energy γ-ray afterglow emission of nearby gamma-ray bursts. *Astrophys. J.* **703**, 60–67 (2009).
27. Piran, T. & Nakar, E. On the external shock synchrotron model for gamma-ray bursts' GeV emission. *Astrophys. J. Lett.* **718**, 63–67 (2010).
28. Tam, P.-H. T., Tang, Q.-W., Hou, S.-J., Liu, R.-Y. & Wang, X.-Y. Discovery of an extra hard spectral component in the high-energy afterglow emission of GRB 130427A. *Astrophys. J. Lett.* **771**, 13 (2013).
29. Liu, R.-Y., Wang, X.-Y. & Wu, X.-F. Interpretation of the unprecedentedly long-lived high-energy emission of GRB 130427A. *Astrophys. J. Lett.* **773**, 20 (2013).
30. Ackermann, M. et al. Fermi-LAT observations of the gamma-ray burst GRB 130427A. *Science* **343**, 42–47 (2014).
31. Wang, X.-Y., Liu, R.-Y., Zhang, H.-M., Xi, S.-Q. & Zhang, B. Synchrotron self-Compton emission from afterglow shocks as the origin of the sub-TeV emission in GRB 180720B and GRB 190114C. *Astrophys. J.* **884**, 117–121 (2019)

MAGIC Collaboration*, P. Veres[1], P. N. Bhat[1,2], M. S. Briggs[1,2], W. H. Cleveland[3], R. Hamburg[1,2], C. M. Hui[4], B. Mailyan[1], R. D. Preece[1,2], O. J. Roberts[3], A. von Kienlin[5], C. A. Wilson-Hodge[4], D. Kocevski[4], M. Arimoto[6], D. Tak[7,8], K. Asano[9], M. Axelsson[10,11], G. Barbiellini[12], E. Bissaldi[13,14], F. Fana Dirirsa[15], R. Gill[16], J. Granot[16], J. McEnery[7,8], N. Omodei[17,18], S. Razzaque[15], F. Piron[19], J. L. Racusin[8], D. J. Thompson[8], S. Campana[20], M. G. Bernardini[20], N. P. M. Kuin[21], M. H. Siegel[22], S. B. Cenko[8,23], P. O'Brien[24], M. Capalbi[25], A. Dai[25], M. De Pasquale[26], J. Gropp[22], N. Klingler[22], J. P. Osborne[24], M. Perri[27,28], R. L. C. Starling[24], G. Tagliaferri[20,25], A. Tohuvavohu[22], A. Ursi[29], M. Tavani[29,30,31], M. Cardillo[29], C. Casentini[29], G. Piano[29], Y. Evangelista[29], F. Verrecchia[27,28], C. Pittori[27,28], F. Lucarelli[27,28], A. Bulgarelli[28], N. Parmiggiani[28], G. E. Anderson[32], J. P. Anderson[33], G. Bernardi[34,35,36], J. Bolmer[5], M. D. Caballero-García[37], I. M. Carrasco[38], A. Castellón[39], N. Castro Segura[40], A. J. Castro-Tirado[41,42], S. V. Cherukuri[43], A. M. Cockeram[44], P. D'Avanzo[20], A. Di Dato[45,46], R. Diretse[47], R. P. Fender[48], E. Fernández-García[42], J. P. U. Fynbo[49,50], A. S. Fruchter[51], J. Greiner[5], M. Gromadzki[52], K. E. Heintz[53], I. Heywood[35,48], A. J. van der Horst[54,55], Y.-D. Hu[42,56], C. Inserra[57], L. Izzo[42,58], V. Jaiswal[43], P. Jakobsson[53], J. Japelj[59], E. Kankare[60], D. A. Kann[42], C. Kouveliotou[54,55], S. Klose[61], A. J. Levan[62], X. Y. Li[63,64], S. Lotti[29], K. Maguire[65], D. B. Malesani[49,50,58,66], I. Manulis[67], M. Marongiu[68,69], S. Martin[33,70], A. Melandri[20], M. J. Michałowski[71], J. C. A. Miller-Jones[32], K. Misra[72,73], A. Moin[74], K. P. Mooley[75,76], S. Nasri[74], M. Nicholl[77,78], A. Noschese[45], G. Novara[79,80], S. B. Pandey[72], E. Peretti[68,81], C. J. Pérez del Pulgar[41], M. A. Pérez-Torres[42,82], D. A. Perley[44], L. Piro[29], F. Ragosta[46,83,84], L. Resmi[43], R. Ricci[34], A. Rossi[85], R. Sánchez-Ramírez[29], J. Selsing[50], S. Schulze[86], S. J. Smartt[87], I. A. Smith[88], V. V. Sokolov[89], J. Stevens[90], N. R. Tanvir[24], C. C. Thöne[42], A. Tiengo[79,80,91], E. Tremou[92], E. Troja[8,93], A. de Ugarte Postigo[42,58], A. F. Valeev[89], S. D. Vergani[94], M. Wieringa[95], P. A. Woudt[47], D. Xu[96], O. Yaron[67] & D. R. Young[87]

MAGIC Collaboration*

V. A. Acciari[97], S. Ansoldi[98,99], L. A. Antonelli[100], A. Arbet Engels[101], D. Baack[102], A. Babić[103], B. Banerjee[104], U. Barres de Almeida[105], J. A. Barrio[106], J. Becerra González[97], W. Bednarek[107], L. Bellizzi[108], E. Bernardini[109,110], A. Berti[111], J. Besenrieder[112], W. Bhattacharyya[109], C. Bigongiari[100], A. Biland[101], O. Blanch[113], G. Bonnoli[108], Ž. Bošnjak[103], G. Busetto[110], R. Carosi[114], G. Ceribella[112], Y. Chai[112], A. Chilingaryan[115], S. Cikota[103], S. M. Colak[113], U. Colin[112], E. Colombo[1], J. L. Contreras[106], J. Cortina[116], S. Covino[100], V. D'Elia[100], P. Da Vela[114], F. Dazzi[100], A. De Angelis[110], B. De Lotto[98], M. Delfino[113,117], J. Delgado[113,117], D. Depaoli[111], F. Di Pierro[111], L. Di Venere[111], E. Do Souto Espiñeira[113], D. Dominis Prester[118], A. Donini[98], D. Dorner[119], M. Doro[110], D. Elsaesser[102], V. Fallah Ramazani[120], A. Fattorini[102], G. Ferrara[100], D. Fidalgo[106], L. Foffano[110], M. V. Fonseca[106], L. Font[121], C. Fruck[112], S. Fukami[122], R. J. García López[1], M. Garczarczyk[109], S. Gasparyan[123], M. Gaug[121], N. Giglietto[111], F. Giordano[111], N. Godinović[124], D. Green[112], D. Guberman[113], D. Hadasch[122], A. Hahn[112], J. Herrera[1], J. Hoang[106], D. Hrupec[125], M. Hütten[112], T. Inada[122], S. Inoue[126], K. Ishio[112], Y. Iwamura[122], L. Jouvin[113], D. Kerszberg[113], H. Kubo[99], J. Kushida[127], A. Lamastra[100], D. Lelas[124], F. Leone[100], E. Lindfors[120], S. Lombardi[100], F. Longo[98,128,129], M. López[106], R. López-Coto[110], A. López-Oramas[1], S. Loporchio[111], B. Machado de Oliveira Fraga[105], C. Maggio[121], P. Majumdar[104], M. Makariev[130], M. Mallamaci[110], G. Maneva[130], M. Manganaro[118], K. Mannheim[119], L. Maraschi[100], M. Mariotti[110], M. Martínez[113], D. Mazin[112,122], S. Mićanović[118], D. Miceli[98], M. Minev[130], J. M. Miranda[108], R. Mirzoyan[112], E. Molina[131], A. Moralejo[113], D. Morcuende[106], V. Moreno[121], E. Moretti[113], P. Munar-Adrover[121], V. Neustroev[132], C. Nigro[109], K. Nilsson[120], D. Ninci[113], K. Nishijima[127], K. Noda[122], L. Nogués[113], S. Nozaki[99], S. Paiano[110], M. Palatiello[98], D. Paneque[112], R. Paoletti[108], J. M. Paredes[131], P. Peñil[106], M. Peresano[98], M. Persic[98], P. G. Prada Moroni[114], E. Prandini[110], I. Puljak[124], W. Rhode[102], M. Ribó[131], J. Rico[113], C. Righi[100], A. Rugliancich[114], L. Saha[106], N. Sahakyan[123], T. Saito[122], S. Sakurai[122], K. Satalecka[109], K. Schmidt[102], T. Schweizer[112], J. Sitarek[107], I. Šnidarić[133], D. Sobczynska[107], A. Somero[1], A. Stamerra[4], D. Strom[112], M. Strzys[112], Y. Suda[112], T. Surić[133], M. Takahashi[122], F. Tavecchio[4], P. Temnikov[130], T. Terzić[118,133], M. Teshima[112,122], N. Torres-Albà[131], L. Tosti[111], V. Vagelli[111], J. van Scherpenberg[112], G. Vanzo[1], M. Vazquez Acosta[1], C. F. Vigorito[111], V. Vitale[111], I. Vovk[112], M. Will[112], D. Zarić[124] & L. Nava[4,129,12]

[1]Center for Space Plasma and Aeronomic Research, University of Alabama in Huntsville, Huntsville, AL, USA. [2]Space Science Department, University of Alabama in Huntsville, Huntsville, AL, USA. [3]Science and Technology Institute, Universities Space Research Association, Huntsville, AL, USA. [4]Astrophysics Branch, ST12, NASA/Marshall Space Flight Center, Huntsville, AL, USA. [5]Max-Planck Institut für extraterrestrische Physik, Garching, Germany. [6]Faculty of Mathematics and Physics, Institute of Science and Engineering, Kanazawa University, Kanazawa, Japan. [7]Department of Physics, University of Maryland, College Park, MD, USA. [8]Astrophysics Science Division, NASA Goddard Space Flight Center, Greenbelt, MD, USA. [9]Institute for Cosmic-Ray Research, University of Tokyo, Kashiwa, Japan. [10]Department of Physics, Stockholm University, Stockholm, Sweden. [11]Department of Physics, KTH Royal Institute of Technology, Stockholm, Sweden. [12]Istituto Nazionale Fisica Nucleare (INFN), Trieste, Italy. [13]Dipartimento di Fisica "M. Merlin" dell'Università e del Politecnico di Bari, Bari, Italy. [14]Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Bari, Italy. [15]Department of Physics, University of Johannesburg, Auckland Park, South Africa. [16]Department of Natural Sciences, Open University of Israel, Ra'anana, Israel. [17]W. W. Hansen Experimental Physics Laboratory, Kavli Institute for Particle Astrophysics and Cosmology, Department of Physics Stanford, Stanford, CA, USA. [18]SLAC National Accelerator Laboratory, Stanford University, Stanford, CA, USA. [19]Laboratoire Univers et Particules de Montpellier, Université Montpellier, CNRS/IN2P3, Montpellier, France. [20]INAF, Astronomical Observatory of Brera, Merate, Italy. [21]Mullard Space Science Laboratory, University College London, Dorking, UK. [22]Department of Astronomy and Astrophysics, Pennsylvania State University, University Park, PA, USA. [23]Joint Space-Science Institute, University of Maryland, College Park, MD, USA. [24]Department of Physics and Astronomy, University of Leicester, Leicester, UK. [25]INAF Istituto di Astrofisica Spaziale e Fisica Cosmica di Palermo, Palermo, Italy. [26]Department of Astronomy and Space Sciences, Istanbul University, Istanbul, Turkey. [27]INAF, Osservatorio Astronomico di Roma, Rome, Italy. [28]Space Science Data Center (SSDC), Agenzia Spaziale Italiana (ASI), Rome, Italy. [29]INAF-IAPS, Rome, Italy. [30]Università "Tor Vergata", Rome, Italy. [31]Gran Sasso Science Institute, L'Aquila, Italy. [32]International Centre for Radio Astronomy Research, Curtin University, Perth, Western Australia, Australia. [33]European Southern Observatory, Santiago, Chile. [34]INAF Istituto di Radioastronomia, Bologna, Italy. [35]Department of Physics and Electronics, Rhodes University, Grahamstown, South Africa. [36]South African Radio Astronomy Observatory, Cape Town, South Africa. [37]Astronomical Institute of the Academy of Sciences, Prague, Czech Republic. [38]Departamento de Física Aplicada, Facultad de Ciencias, Universidad de Málaga, Málaga, Spain. [39]Departamento de Álgebra, Geometría y Topología, Facultad de Ciencias, Universidad de Málaga, Málaga, Spain. [40]Physics and Astronomy Department, University of Southampton, Southampton, UK. [41]Unidad Asociada al CSIC Departamento de Ingeniería de Sistemas y Automática, E.T.S. de Ingenieros Industriales, Universidad de Málaga, Málaga, Spain. [42]Instituto de Astrofísica de Andalucía (IAA-CSIC), Granada, Spain. [43]Indian Institute of Space Science & Technology, Trivandrum, India. [44]Astrophysics Research Institute, Liverpool John Moores University, Liverpool, UK. [45]Osservatorio Astronomico 'S. Di Giacomo' AstroCampania, Agerola, Italy. [46]INAF - Astronomical Observatory of Naples, Naples, Italy. [47]Inter-University Institute for Data-Intensive Astronomy, Department of Astronomy, University of Cape Town, Rondebosch, South Africa. [48]Department of Physics, University of Oxford, Keble Road, Oxford, UK. [49]Cosmic Dawn Center (DAWN), Copenhagen, Denmark. [50]Niels Bohr Institute, Copenhagen University, Copenhagen, Denmark. [51]Space Telescope Science Institute, Baltimore, MD, USA. [52]Astronomical Observatory, University of Warsaw, Warsaw, Poland. [53]Centre for Astrophysics and Cosmology, Science Institute, University of Iceland, Reykjavik, Iceland. [54]Department of Physics, The George Washington University, Washington, DC, USA. [55]Astronomy, Physics, and Statistics Institute of Sciences (APSIS), The George Washington University, Washington, DC, USA. [56]Universidad de Granada, Facultad de Ciencias Campus Fuentenueva, Granada, Spain. [57]School of Physics & Astronomy, Cardiff University, Cardiff, UK. [58]DARK, Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark. [59]Anton Pannekoek Institute for Astronomy, University of Amsterdam, Amsterdam, The Netherlands. [60]Tuorla Observatory, Department of Physics and Astronomy, University of Turku, Turku, Finland. [61]Thüringer Landessternwarte Tautenburg, Tautenburg, Germany. [62]Department of Astrophysics/IMAPP, Radboud University, Nijmegen, The Netherlands. [63]Instituto de Hortofruticultura Subtropical y Mediterránea La Mayora (IHSM/UMA-CSIC), Málaga, Spain. [64]Nanjing Institute for Astronomical Optics and Technology, National Observatories, Chinese Academy of Sciences, Nanjing, China. [65]School of Physics, Trinity College Dublin, Dublin, Ireland. [66]DTU Space, National Space Institute, Technical University of Denmark, Kongens Lyngby, Denmark. [67]Benoziyo Center for Astrophysics, Weizmann Institute of Science, Rehovot, Israel. [68]Department of Physics and Earth Science, University of Ferrara, Ferrara, Italy. [69]International Center for Relativistic Astrophysics Network (ICRANet), Pescara, Italy. [70]Joint ALMA Observatory, Santiago, Chile. [71]Astronomical Observatory Institute, Faculty of Physics, Adam Mickiewicz University, Poznan, Poland. [72]Aryabhatta Research Institute of Observational Sciences, Nainital, India. [73]Department of Physics, University of California, Davis, CA, USA. [74]Physics Department, United Arab Emirates University, Al-Ain, United Arab Emirates. [75]National Radio Astronomy Observatory, Socorro, NM, USA. [76]Caltech, Pasadena, CA, USA. [77]Institute for Astronomy, University of Edinburgh, Royal Observatory, Edinburgh, UK. [78]Birmingham Institute for Gravitational Wave Astronomy and School of Physics and Astronomy, University of Birmingham, Birmingham, UK. [79]Scuola Universitaria Superiore IUSS

Pavia, Pavia, Italy. [80]INAF – IASF Milano, Milan, Italy. [81]INFN, Laboratori Nazionali del Gran Sasso, Assergi, Italy. [82]Departamento de Física Teórica, Universidad de Zaragoza, Zaragoza, Spain. [83]Dipartimento di Scienze Fisiche, Università degli studi di Napoli Federico II, Naples, Italy. [84]INFN Sezione di Napoli, Complesso Universitario di Monte S. Angelo, Naples, Italy. [85]INAF Osservatorio di Astrofisica e Scienza dello Spazio, Bologna, Italy. [86]Department of Particle Physics and Astrophysics, Weizmann Institute of Science, Rehovot, Israel. [87]Astrophysics Research Centre, School of Mathematics and Physics, Queen's University Belfast, Belfast, UK. [88]Department of Physics and Astronomy, Rice University, Houston, TX, USA. [89]Special Astrophysical Observatory (SAO-RAS), Nizhniy Arkhyz, Russia. [90]CSIRO Australia Telescope National Facility, Paul Wild Observatory, Narrabri, New South Wales, Australia. [91]Istituto Nazionale di Fisica Nucleare, Sezione di Pavia, Pavia, Italy. [92]AIM, CEA, CNRS, Université Paris Diderot, Sorbonne Paris Cité, Université Paris-Saclay, Gif-sur-Yvette, France. [93]Department of Astronomy, University of Maryland, College Park, MD, USA. [94]GEPI, Observatoire de Paris, PSL University, CNRS, Meudon, France. [95]Australia Telescope National Facility, CSIRO Astronomy and Space Science, Epping, New South Wales, Australia. [96]CAS Key Laboratory of Space Astronomy and Technology, National Astronomical Observatories, Chinese Academy of Sciences, Beijing, 100012, China. [97]Instituto de Astrofísica de Canarias and Departamento Astrofísica, Universidad de La Laguna, La Laguna, Spain. [98]Università di Udine and INFN Trieste, Udine, Italy. [99]Japanese MAGIC Consortium, Department of Physics, Kyoto University, Kyoto, Japan. [100]National Institute for Astrophysics (INAF), Rome, Italy. [101]ETH Zurich, Zurich, Switzerland. [102]Technische Universität Dortmund, Dortmund, Germany. [103]Croatian Consortium, University of Zagreb, FER, Zagreb, Croatia. [104]Saha Institute of Nuclear Physics, HBNI, Kolkata, India. [105]Centro Brasileiro de Pesquisas Físicas (CBPF), Rio de Janeiro, Brazil. [106]IPARCOS Institute and EMFTEL Department, Universidad Complutense de Madrid, Madrid, Spain. [107]University of Łódź, Department of Astrophysics, Łódź, Poland. [108]Università di Siena and INFN Pisa, Siena, Italy. [109]Deutsches Elektronen-Synchrotron (DESY), Zeuthen, Germany. [110]Università di Padova and INFN, Padua, Italy. [111]Istituto Nazionale Fisica Nucleare (INFN), Frascati, Italy. [112]Max-Planck-Institut für Physik, Munich, Germany. [113]Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [114]Università di Pisa and INFN Pisa, Pisa, Italy. [115]The Armenian Consortium, A. Alikhanyan National Laboratory, Yerevan, Armenia. [116]Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas, Madrid, Spain. [117]Port d'Informació Científica (PIC), Barcelona, Spain. [118]Croatian Consortium, Department of Physics, University of Rijeka, Rijeka, Croatia. [119]Universität Würzburg, Würzburg, Germany. [120]Finnish MAGIC Consortium, Finnish Centre of Astronomy with ESO (FINCA), University of Turku, Turku, Finland. [121]Departament de Física and CERES-IEEC, Universitat Autònoma de Barcelona, Bellaterra, Spain. [122]Japanese MAGIC Consortium, ICRR, The University of Tokyo, Kashiwa, Japan. [123]The Armenian Consortium, ICRANet-Armenia at NAS RA, Yerevan, Armenia. [124]Croatian Consortium, University of Split, FESB, Split, Croatia. [125]Croatian Consortium, Josip Juraj Strossmayer University of Osijek, Osijek, Croatia. [126]Japanese MAGIC Consortium, RIKEN, Wako, Japan. [127]Japanese MAGIC Consortium, Tokai University, Hiratsuka, Japan. [128]Dipartimento di Fisica, Università di Trieste, Trieste, Italy. [129]Institute for Fundamental Physics of the Universe (IFPU), Trieste, Italy. [130]Institute for Nuclear Research and Nuclear Energy, Bulgarian Academy of Sciences, Sofia, Bulgaria. [131]Universitat de Barcelona, ICCUB, IEEC-UB, Barcelona, Spain. [132]Finnish MAGIC Consortium, Astronomy Research Unit, University of Oulu, Oulu, Finland. [133]Croatian Consortium, Rudjer Boskovic Institute, Zagreb, Croatia . *e-mail: contact.magic@mpp.mpg.de

Methods

Prompt-emission observations

On 14 January 2019, the prompt emission from GRB 190114C triggered several space instruments, including Fermi-GBM[32], Fermi-LAT[33], Swift-BAT[34], Super-AGILE[35], AGILE-MCAL[35], KONUS-Wind[36], INTEGRAL-SPI-ACS[37] and Insight-HXMT[38]. The prompt-emission light curves from AGILE, Fermi and Swift are shown in Fig. 1 and in Extended Data Fig. 1, where the trigger time T_0 refers to the BAT trigger time (20:57:03.19 UT). The prompt emission lasts for approximately 25 s, when the last flaring-emission episode ends. Nominally, T_{90} (that is, the time interval during which a fraction between 5% and 95% of the total emission is observed) is much longer (>100 s, depending on the instrument)[14], but it is clearly contaminated by the afterglow component (Fig. 1) and does not provide a good measure of the actual duration of the prompt emission. A more detailed study of the prompt emission phase is reported in ref. [14].

AGILE

AGILE (Astrorivelatore Gamma ad Immagini Leggero)[39] could observe GRB 190114C until $T_0 + 330$ s, before it became occulted by the Earth. GRB 190114C triggered the MCAL (Mini-CALorimeter) from $T_0 - 0.95$ s to $T_0 + 10.95$ s. The MCAL light-flux curve in Fig. 1 was produced using two different spectral models. From $T_0 - 0.95$ s to $T_0 + 1.8$ s, the spectrum is fitted by a power law with photon index $\Gamma_{\rm ph} = -1.97^{+0.47}_{-0.70}$ ($\mathrm{d}N/\mathrm{d}E \propto E^{\Gamma_{\rm ph}}$). From $T_0 + 1.8$ s to $T_0 + 5.5$ s the best-fit model is a broken power law with $\Gamma_{\rm ph,1} = -1.87^{+0.54}_{-0.19}$, $\Gamma_{\rm ph,2} = -2.63^{+0.07}_{-0.07}$ and break energy $E_{\rm b} = 756^{+137}_{-159}$ keV. The total fluence in the 0.4–100 MeV energy range is $F = 1.75 \times 10^{-4}$ erg cm^{-2}. The Super-AGILE detector also detected the burst, but the large off-axis angle prevented any X-ray imaging of the burst and any spectral analysis. Extended Data Fig. 1a, d, e shows the GRB 190114C light curves acquired by the Super-AGILE detector (20–60 keV) and by the MCAL detector in the low- (0.4–1.4 MeV) and high-energy (1.4–100 MeV) bands.

Fermi-GBM

There are indications that at the time of the MAGIC observations some of the detectors were partially shadowed by the structural elements of the Fermi spacecraft that were not modelled in the response of the GBM (Gamma-ray Burst Monitor) detectors. This affects the low-energy part of the spectrum[40]. For this reason, out of caution we elected to exclude the energy channels below 50 keV. The spectra detected by Fermi-GBM[41] during the intervals $T_0 + 68$ s to $T_0 + 110$ s and $T_0 + 110$ s to $T_0 + 180$ s are best described by a power-law model with photon index $\Gamma_{\rm ph} = -2.10 \pm 0.08$ and $\Gamma_{\rm ph} = -2.05 \pm 0.10$, respectively (Figs. 2, 3). The 10–1,000-keV light curve in Extended Data Fig. 1c was constructed by summing photon counts for the bright NaI detectors.

Swift-BAT

The 15–350-keV mask-weighted light curve of the BAT (Burst Alert Telescope)[42] shows a multi-peaked structure that starts at $T_0 - 7$ s (Extended Data Fig. 1b). The 68–110 s and 110–180 s spectra shown in Figs. 2, 3 were derived from a joint XRT–BAT fit. The best-fitting parameters for the whole interval (68–180 s) are: column density, $N_{\rm H} = (7.53^{+0.74}_{-1.74}) \times 10^{22}$ cm^{-2} at $z = 0.42$, in addition to the galactic value of 7.5×10^{19} cm^{-2}; low-energy photon index, $\Gamma_{\rm ph,1} = -1.21^{+0.40}_{-1.26}$; high-energy spectral index, $\Gamma_{\rm ph,2} = -2.19^{+0.39}_{-0.19}$; and peak energy $E_{\rm pk} > 14.5$ keV. Errors are given at 90% confidence level.

Fermi-LAT

Fermi-LAT (Large Area Telescope)[43] detected a γ-ray counterpart since the prompt phase[33]. The burst left the LAT field of view at $T_0 + 150$ s and remained outside it until $T_0 + 8{,}600$ s. The light curve in the energy range 0.1–10 GeV is shown in Extended Data Fig. 1f. The LAT spectra in the time bins 68–110 s and 110–180 s (Figs. 2, 3) are described by a power law with pivot energies of 200 MeV and 500 MeV, photon indices $\Gamma_{\rm ph}(68\text{–}110) = -2.02 \pm 0.95$ and $\Gamma_{\rm ph}(110\text{–}180) = -1.69 \pm 0.42$, and normalization factors of $N_{0,68\text{–}110} = (2.02 \pm 1.31) \times 10^{-7}$ MeV^{-1} cm^{-2} s^{-1} and $N_{0,110\text{–}180} = (4.48 \pm 2.10) \times 10^{-8}$ MeV^{-1} cm^{-2} s^{-1}, respectively. In each time interval, the analysis was limited to the energy range in which photons were detected. The LAT light curve integrated in the energy range 0.1–1 GeV is shown in Fig. 1.

MAGIC

To analyse the data we used the standard MAGIC software[44] and followed the steps optimized for data taking under moderate moon illumination[45]. The spectral fitting was performed by a forward-folding method, assuming a simple power law for the intrinsic spectrum and taking into account the EBL effect, using the model of Domínguez et al.[46]. Extended Data Table 1 shows the fitting results for various time bins (the pivot energy is chosen to minimize the correlation between the normalization and photon index parameters). The data points shown in Figs. 2, 3 were obtained from the observed excess rates in estimated energy, the fluxes of which were evaluated in true energy (photon corrected energy by Monte Carlo simulation, after reconstruction and unfolding) using the effective time and a spill-over-corrected effective area obtained from the best fit.

The time-resolved analysis hints to a possible spectral evolution to softer values, although we cannot exclude that the photon indices are compatible with a constant value of about −2.5 up to 2,400 s. The signal and background in the considered time bins are both in the low-count Poisson regime. Therefore, the correct treatment of the MAGIC data provided here includes the use of Poisson statistics, as well as systematic errors. To estimate the main source of systematic errors—our imperfect knowledge of the absolute instrument calibration and the total atmospheric transmission—we vary the light scale in our Monte Carlo simulation, as suggested in previous studies[44]. The result is reported in the last two lines of Extended Data Table 1 and in Extended Data Fig. 2.

The systematic effects deriving from the choice of one particular EBL model were also studied. The analysis performed to obtain the time-integrated spectrum was repeated, employing three other models[47–49]. The contribution to the systematic error on the photon index caused by the uncertainty on the EBL model is $\sigma_\alpha = ^{+0.10}_{-0.13}$, which is smaller than the statistical error only (one standard deviation), as already seen in a previous work[10]. On the other hand, the contribution of the choice of the EBL model to the systematic error on the normalization factor is only partially at the same level of the statistical error (one standard deviation), $\sigma_{\rm N} = ^{+0.30}_{-0.08} \times 10^{-8}$. The chosen EBL model returns a normalization factor that is lower than two of the other models and very close to the third one[47].

The MAGIC energy-flux light curve that is presented in Fig. 1 was obtained by integrating the best-fit spectral model of each time bin from 0.3 to 1 TeV, in the same manner as in a previous study[10]. The value of the fitted time constant reported here differs less than two standard deviations from the one previously reported[10]. The difference is due to the poor constraints on the spectral-fit parameters of the last time bin, which influences the light-curve fit.

X-ray afterglow observations

Swift/XRT. Swift-XRT (X-Ray Telescope) started observing 68 s after T_0. The source light curve[50] was taken from the Swift-XRT light-curve repository[51] and was converted into 1–10-keV flux (Fig. 1) through dedicated spectral fits. The combined XRT + BAT spectral fit in Figs. 2, 3 is described above.

XMM-Newton and NuSTAR. The XMM-Newton X-ray observatory and the Nuclear Spectroscopic Telescope Array (NuSTAR) started observing GRB 190114C under Director's Discretionary Time (DDT) Target of Opportunities 7.5 h and 22.5 h, respectively, after the burst. The XMM-Newton and NuSTAR absorption-corrected fluxes (Fig. 1) were derived by fitting the spectrum with XSPEC and with the same

power-law model, considering absorption in our Galaxy and at the redshift of the burst.

NIR, optical and UV afterglow observations

Light curves from the different instruments presented in this section are shown in Extended Data Fig. 3.

GROND. The Gamma-Ray burst Optical/Near-infrared Detector (GROND)[52] started observations 3.8 h after the GRB trigger, and the follow-up continued until 29 January 29 2019. Image reduction and photometry were carried out with standard IRAF tasks[53], as described in refs. [54,55]. JHK_s photometry was converted to AB magnitudes to provide a common flux system. The final photometry is given in Extended Data Table 2.

BOOTES and GTC. The CASANDRA-1 ultra-wide-field camera[56] at the BOOTES-1 station in ESAt/INTA-CEDEA (Huelva, Spain) took an image of the GRB 190114C location, starting at 20:57:18 UT (30 s exposure time) (Extended Data Fig. 4). The Gran Canarias Telescope (GTC), equipped with the OSIRIS spectrograph[57], started observations 2.6 h post-burst. The grisms R1000B and R2500I were used, covering the wavelength range 3,700–10,000 Å (600 s exposure time for each grism). The GTC detected a highly extinguished continuum, as well as Ca II H and K lines in absorption and [O II], H_β and [O III] in emission (see Extended Data Fig. 5), all roughly at the same redshift of $z = 0.4245 \pm 0.0005$ (ref. [58]). By comparing the derived rest-frame equivalent widths with ref. [59], GRB 190114C clearly shows higher than average, but not unprecedented, values.

HST. The Hubble Space Telescope (HST) imaged the afterglow and host galaxy of GRB 190114C on 11 February and 12 March 2019. HST observations clearly reveal that the host galaxy is spiral (Extended Data Fig. 4). A direct subtraction of the epochs of observations with the F850LP filter yields a faint residual close to the nucleus of the host (Extended Data Fig. 4). From the position of the residual we estimate that the burst originated within 250 pc of the host galaxy nucleus.

LT. The robotic 2-m Liverpool Telescope (LT)[60] slewed to the afterglow location at coordinated universal time (UTC) 2019-01-14 23:22:34 and on the second night from UTC 2019-01-15 19:32:10 and acquired images in the B, g, V, r, i and z bands (45 s exposure each on the first night and 60 s on the second; see Extended Data Table 3). Aperture photometry of the afterglow was performed using a custom IDL script with a fixed aperture radius of 1.5″. Photometric calibration was performed relative to stars from the Pan-STARRS1 catalogue[61].

NTT. The European Southern Observatory's (ESO) New Technology Telescope (NTT) observed the optical counterpart of GRB 190114C under the extended Public ESO Spectroscopic Survey for Transient Objects (ePESSTO) using the NTT/EFOSC2 instrument in imaging mode[62]. Observations started at 04:36:53 UT on 16 January 2019 with g, r, i and z Gunn filters. Image reduction was carried out by following the standard procedures[63].

OASDG. The 0.5-m remote telescope of the Osservatorio Astronomico 'S. Di Giacomo' (OASDG), located in Agerola (Italy), started observations in the optical RC band 0.54 h after the burst. The afterglow of GRB 190114C was clearly detected in all the images.

NOT. The Nordic Optical Telescope (NOT) observed the optical afterglow of GRB 190114C with the Alhambra Faint Object Spectrograph and Camera (AlFOSC) instrument. Imaging was obtained in the griz filters with 300-s exposures, starting at 14 January 2019 21:20:56 UT, 24 min after the BAT trigger. The normalized spectrum (Extended Data Fig. 5) reveals strong host interstellar absorption lines of Ca H and K and of Na I D, which provided a redshift of $z = 0.425$.

REM. The 60-cm robotic Rapid Eye Mount telescope (REM) performed optical and NIR observations with the ROS2 optical imager and the REMIR NIR camera[64]. Observations were performed starting about 3.8 h after the burst in the r and J bands and lasted about one hour.

Swift/UVOT. The Swift UltraViolet and Optical Telescope (UVOT)[65] began observations at $T_0 + 54$ s in the UVOT v-band. The first observation after settling was in the UVOT white band[66], started 74 s after the trigger and lasted for 150 s. A 50-s exposure with the UV grism was taken next, followed by multiple exposures rotating through all seven broad- and intermediate-band filters, until switching to only the UVOT clear white filter on 20 January 2019. Standard photometric calibration and methods were used to derive the aperture photometry[67,68]. The grism zeroth-order data were reduced manually[69] to derive the B-magnitude and error.

VLT. The STARGATE collaboration used the Very Large Telescope (VLT) and observed GRB 190114C using the X-shooter spectrograph. Detailed analysis will be presented in forthcoming papers. A portion of the second spectrum is shown in Extended Data Fig. 5, illustrating the strong emission lines that are characteristic of a strongly star-forming galaxy, whose light is largely dominating over the afterglow at this epoch.

Magnitudes of the underlying galaxies

The HST images show a spiral or tidally disrupted galaxy whose bulge is coincident with the coordinates of GRB 190114C. A second galaxy is detected at an angular distance of 1.3″ towards the northeast. The SED analysis was performed with LePhare[70,71] using an iterative method that combined both the resolved photometry of the two galaxies found in the HST and VLT/HAWK-I data and the blended photometry from GALEX and WISE, in which the spatial resolution was much lower. Further details will be given in a separate paper (A.d.U.P. et al., manuscript in preparation). The estimated photometry for each object and their combination is given in Extended Data Table 4.

Optical extinction

The optical extinction towards the line of sight of a GRB is derived assuming a power law as the intrinsic spectral shape[72]. Once the Galactic extinction ($E_{B-V} = 0.01$; ref. [73]) is taken into account and the fairly bright host galaxy contribution is properly subtracted, a good fit to the data is obtained with the Large Magellanic Cloud recipe and $A_V = 1.83 \pm 0.15$. The spectral index β ($F_\nu \propto \nu^{\beta_o}$) evolves from hard to soft across the temporal break in the optical light curve at about 0.5 d, moving from $\beta_{o,1} = -0.10 \pm 0.12$ to $\beta_{o,2} = -0.48 \pm 0.15$.

Radio and submillimetre afterglow observations

The light curves obtained by the different instruments are shown in Extended Data Fig. 3.

ALMA. Observations with the Atacama Large Millimetre–Submillimetre Array (ALMA) are reported in Band 3 (central observed frequency of 97.500 GHz) and Band 6 (235.0487 GHz) between 15 January and 19 January 2019. The data were calibrated within CASA (Common Astronomy Software Applications; version 5.4.0)[74] using the pipeline calibration. Photometric measurements were also performed within CASA. Early ALMA observations at 97.5 GHz are taken from ref. [16].

ATCA. The Australia Telescope Compact Array (ATCA) observations were made with the ATCA 4-cm receivers (band centres of 5.5 and 9 GHz), 15-mm receivers (band centres of 17 and 19 GHz) and 7-mm receivers (band centres of 43 and 45 GHz). The ATCA data (see Extended Data Table 5) were obtained using the CABB continuum mode[75] and were reduced with the software packages Miriad[76] and CASA[74] using standard techniques. The quoted errors are 1σ, which include the root-mean-square (r.m.s.) and Gaussian 1σ errors.

GMRT. The upgraded Giant Metre-wave Radio Telescope[77] (UGMRT) observed on 17 January 2019 13.44 UT (2.8 d after the burst) in band 5 (1,000–1,450 MHz) with 2,048 channels spread over 400 MHz. The GMRT detected a weak source with a flux density of 73 ± 17 μJy at the GRB position[78]. The flux should be considered as an upper limit, as the contribution from the host[79] has not been subtracted.

MeerKAT. The new MeerKAT radio observatory[80,81] observed on 15 and 18 January 2019, with DDT requested by the ThunderKAT Large Survey Project[82]. Both epoch measurements used 63 antennas and were carried out in the L-band, spanning 856 MHz and centred at 1,284 MHz. The MeerKAT flux estimation was done by finding and fitting the source with the software PyBDSF v.1.8.15[83]. Adding the r.m.s. noise in quadrature to the flux uncertainty leads to final flux measurements of 125 ± 14 μJy per beam on 15 January and 97 ± 16 μJy per beam on 18 January. The contribution from the host galaxy[79] has not been subtracted. Therefore, these measurements provide a maximum flux of the GRB.

JCMT SCUBA-2. Sub-millimetre observations (Extended Data Table 5) were performed simultaneously at 850 μm and 450 μm on three nights using the Submillimetre Common-User Bolometer Array 2 (SCUBA-2) continuum camera[84] on the James Clerk Maxwell Telescope (JCMT). GRB 190114C was not detected on any of the individual measurements. By combining all the SCUBA-2 continuum camera[84] observations, the r.m.s. background noise is 0.95 mJy per beam at 850 μm and 5.4 mJy per beam at 450 μm at 1.67 d after the burst trigger.

Prompt-emission model for the early-time MAGIC emission

In the standard picture the prompt sub-megaelectronvolt spectrum is explained as synchrotron radiation from relativistic accelerated electrons in the energy-dissipation region. The associated inverse Compton component is sensitive to the details of the dynamics: for example, in the internal shock model if the peak energy is initially very high and the inverse Compton component is suppressed owing to Klein–Nishina effects, the peak of the inverse Compton component may be delayed and become bright only at late times, when scattering occurs in the Thomson regime. Simulations showed that the magnetic fields required to produce the gigaelectronvolt–teraelectronvolt component are rather low[85], with $\varepsilon_B \approx 10^{-3}$. In this framework the contribution of the inverse Compton component to the observed flux at early times (62–90 s; see Extended Data Table 1) does not exceed ~20%. Alternatively, if the prompt emission originates in reprocessed photospheric emission, the early teraelectronvolt flux may arise from inverse Compton scattering of thermal photons by freshly heated electrons below the photosphere at low optical depths. Another possibility for the generation of teraelectronvolt photons might be the inverse Compton scattering of prompt megaelectronvolt photons by electrons in the external forward-shock region, where electrons are heated to an average Lorentz factor of order 10^4 at early times.

Afterglow model

Synchrotron and SSC radiation from electrons accelerated at the forward shock were modelled within the external-shock scenario[7,8,20,25,86]. The results of the modelling are overlaid with the data in Fig. 3 and Extended Data Figs. 6, 7.

We consider two types of power-law radial profiles $n(R) = n_0 R^{-s}$ for the external environment: $s = 0$ (homogeneous medium) and $s = 2$ (wind-like medium, typical of an environment shaped by the stellar wind of the progenitor). In the latter case, we define $n_0 = 3 \times 10^{35} A_* \,\mathrm{cm}^{-1}$, where A_* is a parameter characterizing the normalization of the density. We assume that electrons swept up by the shock are accelerated into a power-law distribution described by the spectral index p, where $\mathrm{d}N/\mathrm{d}\gamma \propto \gamma^{-p}$, where γ is the electron Lorentz factor. We call ν_m the characteristic synchrotron frequency of electrons with Lorentz factor γ_m, ν_c is the cooling frequency and ν_{sa} the synchrotron self-absorption frequency.

The early-time optical emission (up to ~1,000 s) and radio emission (up to ~10^5 s) are probably dominated by reverse-shock radiation[16]. The detailed modelling of this component is not discussed here, where we focus on forward-shock radiation.

The XRT flux (Fig. 1, blue data points) decays as $F_X \propto t^{\alpha_X}$ with $\alpha_X = -1.36 \pm 0.02$. If $\nu_X > \max(\nu_m, \nu_c)$, the X-ray light curve is predicted to decay as $t^{(2-3p)/4}$, which implies $p \approx 2.5$. Another possibility is to assume $\nu_m < \nu_X < \nu_c$, which implies $p = 2.1$–2.2 for $s = 2$ and $p \approx 2.8$ for $s = 0$. A broken power law provides a better fit (5.3×10^{-5} probability of chance improvement), with a break occurring around 4×10^4 s and decay indices of $\alpha_{X,1} \approx -1.32 \pm 0.03$ and $\alpha_{X,2} \approx -1.55 \pm 0.04$. This behaviour can be explained by the passage of ν_c in the XRT band and assuming again $p = 2.4$–2.5 for $s = 2$ and $p \approx 2.8$ for $s = 0$.

The optical light curve starts displaying a shallow decay with time (with temporal index poorly constrained, between −0.5 and −0.25) starting from ~2×10^3 s, followed by a steepening around 8×10^4 s, when the temporal decay becomes similar to the decay in the X-ray band, which suggests that after this time the X-ray and optical bands lie in the same part of the synchrotron spectrum. If the break is interpreted as the synchrotron characteristic frequency ν_m crossing the optical band, after the break the observed temporal decay requires a steep value of $p \approx 3$ for $s = 0$ and a value between $p = 2.4$ and $p = 2.5$ for $s = 2$. Independently of the density profile of the external medium and of the cooling regime of the electrons, $\nu_m \propto t^{-3/2}$, which implies that ν_m is in the soft-X-ray band at 10^2 s. The SED at ~100 s is indeed characterized by a peak between 5–30 keV (Fig. 3). Information on the location of the self-absorption frequency is provided by observations at 1 GHz, showing that $\nu_{sa} \approx 1$ GHz at 10^5 s (Extended Data Fig. 6).

To summarize, in a wind-like scenario, X-ray and optical emission and their evolution in time can be explained if $p = 2.4$–2.5 and the emission is initially in the fast-cooling regime transitions to a slow-cooling regime around 3×10^3 s. The optical spectral index at late times is predicted to be $(1-p)/2 \approx -0.72$, in agreement with observations. ν_m crosses the optical band at $t \approx 8 \times 10^4$ s, explaining the steepening of the optical light curve and the flattening of the optical spectrum. The X-ray band initially lies above (or close to) ν_m, and the break frequency ν_c starts crossing the X-ray band around $(2$–$4) \times 10^4$ s, producing the steepening in the decay rate (the cooling frequency increases with time for $s = 2$). In this case, before the temporal break, the decay rate is related to the spectral index of the electron energy distribution by $\alpha_{X,1} = (2 - 3p)/4 \approx -1.3$ for $p \approx 2.4$–2.5. Well after the break, this value of p predicts a decay rate of $\alpha_{X,1} = (1 - 3p)/4$ between $\alpha_{X,1} = -1.55$ and $\alpha_{X,1} = -1.62$. Overall, this interpretation is also consistent with the fact that the late-time ($t > 10^5$ s) X-ray and optical light curves display similar temporal decays (Fig. 1), as they lie in the same part of the synchrotron spectrum ($\nu_m < \nu_{opt} < \nu_X < \nu_c$). A similar picture can be invoked to explain the emission when assuming a homogeneous density medium, but a steeper value of p is required. In this case, however, no break is predicted in the X-ray light curve.

We now add to the picture the information brought by the teraelectronvolt detection. The model is built with reference to the MAGIC flux and spectral indices derived considering statistical errors only (see Extended Data Table 1 and green data points in Extended Data Fig. 2). The light curve decays in time as $t^{-1.51}$ and the photon index is consistent within ~1σ with $\Gamma_{ph,TeV} \approx -2.5$ for the entire duration of the emission, although there is evidence for an evolution from stronger (about −2) to weaker (about −2.8) values. In the first broadband SED (Fig. 3, 68–110 s), LAT observations provide strong evidence for the presence of two separated spectral peaks.

Assuming Thomson scattering, the SSC peak is given by:

$$\nu_{peak}^{SSC} \approx 2\gamma_e^2 \nu_{peak}^{syn} \quad (1)$$

whereas in the Klein–Nishina regime, the SSC peak should be located at:

$$h\nu_{\rm peak}^{\rm SSC} \approx \frac{2\gamma_{\rm e}\Gamma m_{\rm e}c^2}{1+z} \quad (2)$$

where $\gamma_{\rm e} = \min(\gamma_{\rm c}, \gamma_{\rm m})$. The synchrotron spectral peak is located at $E_{\rm peak}^{\rm syn} \approx 10$ keV and the peak of the SSC component must be $E_{\rm peak}^{\rm SSC} \lesssim 100$ GeV to explain the MAGIC photon index. Both the Klein–Nishina and Thomson scattering regimes imply that $\gamma_{\rm e} \lesssim 10^3$. This small value presents two problems: (i) if the bulk Lorentz factor Γ is larger than 150 (which is a necessary condition to avoid strong γ–γ opacity; see below), a small $\gamma_{\rm m}$ translates into a small efficiency of the electron acceleration, with $\varepsilon_{\rm e} < 0.05$; (ii) the synchrotron peak energy can be located at $E_{\rm peak}^{\rm syn} \approx 10$ keV only for $B\Gamma \gtrsim 10^5$ G. A large B and a small $\varepsilon_{\rm e}$ would make it difficult to explain the presence of a strong SSC emission. These calculations show that γ–γ opacity probably plays a role in shaping and softening the observed SSC spectra[31,87].

For a γ-ray photon with energy E_γ, the $\tau_{\gamma\gamma}$ opacity is:

$$\tau_{\gamma\gamma}(E_\gamma) = \sigma_{\gamma\gamma}(R/\Gamma)n_{\rm t}(E_\gamma) \quad (3)$$

where $n_{\rm t} = L_{\rm t}/(4\pi R^2 c\Gamma E_{\rm t})$ is the density of target photons in the comoving frame, $L_{\rm t}$ is the luminosity and $E_{\rm t} = (m_{\rm e}c^2)^2\Gamma^2/[E_\gamma(1+z)^2]$ is the energy of target photons in the observer frame (c, speed of light in vacuum). Target photons for photons of energy $E_\gamma = 0.2$–1 TeV and for $\Gamma \approx 120$–150 have energies in the range 4–30 keV. When γ–γ absorption is relevant, the emission from pairs can give a non-negligible contribution to the radiative output.

To properly model all the physical processes that shape the broadband radiation, we use a numerical code that solves the evolution of the electron distributions and derives the radiative output, taking into account the following processes: synchrotron and SSC losses, adiabatic losses, γ–γ absorption, emission from pairs and synchrotron self-absorption[88–91]. We find that for the parameters assumed in the proposed model (see below), the contribution from pairs to the emission is negligible.

The MAGIC photon index (Extended Data Table 1) and its evolution with time constrain the SSC peak energy to ≲1 TeV at the beginning of the observations (Extended Data Table 1). In general, the internal opacity decreases with time and Klein–Nishina effects become less relevant. A possible softening of the spectrum with time, as the one suggested by the observations, requires that the spectral peak decreases with time and moves below the MAGIC energy range. In the slow-cooling regime, the SSC peak evolves to higher frequencies for a wind-like medium and decreases very slowly ($\nu_{\rm peak}^{\rm SSC} \propto t^{-1/4}$) for a constant-density medium (both in the Klein–Nishina and Thomson regimes). In the fast-cooling regime the evolution is faster ($\nu_{\rm peak}^{\rm SSC} \propto t^{-1/2} - t^{-9/4}$, depending on the medium and regime).

We model the multi-band observations considering both $s = 0$ and $s = 2$. The results are shown in Fig. 3, Extended Data Figs. 6, 7, where model curves are overlaid with observations. The model curves shown in these figures are derived using the following parameters. For the model in Fig. 3 and in Extended Data Figs. 7 (solid and dashed curves): $s = 0$, $\varepsilon_{\rm e} = 0.07$, $\varepsilon_{\rm B} = 8 \times 10^{-5}$, $p = 2.6$, $n_0 = 0.5$ and $E_{\rm k} = 8 \times 10^{53}$ erg. For the dotted curves in Extended Data Fig. 7 and the SEDs in Extended Data Fig. 6: $s = 2$, $\varepsilon_{\rm e} = 0.6$, $\varepsilon_{\rm B} = 10^{-4}$, $p = 2.4$, $A_* = 0.1$ and $E_{\rm k} = 4 \times 10^{53}$ erg.

Using the constraints on the afterglow onset time ($t_{\rm peak}^{\rm aft} \approx 5 - 10$ s, from the smooth component detected during the prompt emission) the initial bulk Lorentz factor is constrained to values $\Gamma_0 \approx 300$ and $\Gamma_0 \approx 700$ for $s = 2$ and $s = 0$, respectively.

Consistently with the qualitative description above, we find that late-time optical observations can indeed be explained with $\nu_{\rm m}$ crossing the band (see the SED modelling in Extended Data Fig. 6 and the dotted curves in Extended Data Fig. 7). However, a large $\nu_{\rm m}$ is required in this case and consequently the peak of the SSC component would also be large and lie above the MAGIC energy range. The resulting MAGIC light curve (green dotted curve in Extended Data Fig. 7) does not agree with observations. By relaxing the requirement on $\nu_{\rm m}$, the teraelectronvolt spectra (Fig. 3) and light curve (green solid curve in Extended Data Fig. 7) can be explained. As noted, a wind-like medium can explain the steepening of the X-ray light curve at 8×10^4 s, whereas no steepening is expected in a homogeneous medium (blue dotted and solid lines in Extended Data Fig. 7). We find that the gigaelectronvolt flux detected by LAT at a late time ($t \approx 10^4$ s) is dominated by the SSC component (dashed line in Extended Data Fig. 7).

Data availability

Data are available from the corresponding authors upon request.

Code availability

Proprietary data reconstruction codes were generated at the MAGIC telescope large-scale facility. Information supporting the findings of this study is available from the corresponding authors upon request. Source data for Figs. 2, 3 are provided with the paper.

32. Hamburg, R. GRB 190114C: Fermi GBM detection. *GCN Circulars* 23707 https://gcn.gsfc.nasa.gov/gcn3/23707.gcn3 (2019).
33. Kocevski, D. et al. GRB 190114C: Fermi-LAT detection. *GCN Circulars* 23709 https://gcn.gsfc.nasa.gov/gcn3/23709.gcn3 (2019).
34. Gropp, J. D. GRB 190114C: Swift detection of a very bright burst with a bright optical counterpart. *GCN Circulars* 23688 https://gcn.gsfc.nasa.gov/gcn3/23688.gcn3 (2019).
35. Ursi, A. et al. GRB 190114C: AGILE/MCAL detection. *GCN Circulars* 23712 https://gcn.gsfc.nasa.gov/gcn3/23712.gcn3 (2019).
36. Frederiks, D. et al. Konus-Wind observation of GRB 190114C. *GCN Circulars* 23737 https://gcn.gsfc.nasa.gov/gcn3/23737.gcn3 (2019).
37. Minaev, P. & Pozanenko, A. GRB 190114C: SPI-ACS/INTEGRAL extended emission detection. *GCN Circulars* 23714 https://gcn.gsfc.nasa.gov/gcn3/23714.gcn3 (2019).
38. Xiao, S. et al. GRB 190114C: Insight-HXMT/HE detection. *GCN Circulars* 23716 https://gcn.gsfc.nasa.gov/gcn3/23716.gcn3 (2019).
39. Tavani, M. et al. The AGILE mission. *Astron. Astrophys.* **502**, 995–1013 (2009).
40. Goldstein, A. et al. The Fermi GBM gamma-ray burst spectral catalog: the first two years. *Astrophys. J. Suppl. Ser.* **199**, 19 (2012).
41. Meegan, C. et al. The Fermi Gamma-ray Burst Monitor. *Astrophys. J.* **702**, 791–804 (2009).
42. Barthelmy, S. D. et al. The Burst Alert Telescope (BAT) on the SWIFT Midex Mission. *Space Sci. Rev.* **120**, 143–164 (2005).
43. Atwood, A. A. et al. The Large Area Telescope on the Fermi gamma-ray space telescope mission. *Astrophys. J.* **697**, 1071–1102 (2009).
44. Aleksić, J. et al. The major upgrade of the MAGIC telescopes, part II: a performance study using observations of the Crab Nebula. *Astropart. Phys.* **72**, 76–94 (2016).
45. Ahnen, M. L. et al. Performance of the MAGIC telescopes under moonlight. *Astropart. Phys.* **94**, 29–41 (2017).
46. Domínguez, A. et al. Extragalactic background light inferred from AEGIS galaxy-SED-type fractions. *Mon. Not. R. Astron. Soc.* **410**, 2556–2578 (2011).
47. Franceschini, A., Rodighiero, G. & Vaccari, M. Extragalactic optical-infrared background radiation, its time evolution and the cosmic photon-photon opacity. *Astron. Astrophys.* **487**, 837–852 (2008).
48. Finke, J. D., Razzaque, S. & Dermer, C. D. Modeling the extragalactic background light from stars and dust. *Astrophys. J.* **712**, 238–249 (2010).
49. Gilmore, R. C., Somerville, R. S., Primack, J. R. & Domínguez, A. Semi-analytic modelling of the extragalactic background light and consequences for extragalactic gamma-ray spectra. *Mon. Not. R. Astron. Soc.* **422**, 3189–3207 (2012).
50. UK Swift Science Data Centre. GRB 190114C Swift-XRT light curve https://www.swift.ac.uk/xrt_curves/00883832/.
51. Evans, P. A. et al. Methods and results of an automatic analysis of a complete sample of Swift-XRT observations of GRBs. *Mon. Not. R. Astron. Soc.* **397**, 1177–1201 (2009).
52. Greiner, J. et al. GROND—a 7-channel imager. *Publ. Astron. Soc. Pacif.* **120**, 405–424 (2008).
53. Tody, D. in *Astronomical Data Analysis Software and Systems II, ASP Conference Series* Vol. 52 (eds Hanisch, R. J. et al.) 173–183 (1993).
54. Krühler, T. et al. The 2175 Å dust feature in a gamma-ray burst afterglow at redshift 2.45. *Astrophys. J.* **685**, 376–383 (2008).
55. Bolmer, J. et al. Dust reddening and extinction curves toward gamma-ray bursts at $z > 4$. *Astron. Astrophys.* **609**, A62 (2018).
56. Castro-Tirado, A. J. et al. A very sensitive all-sky CCD camera for continuous recording of the night sky. In *Proc. SPIE, Advanced Software and Control for Astronomy II* Vol. 7019 (SPIE, 2008).
57. Cepa, J. et al. OSIRIS tunable imager and spectrograph. In In *Proc. SPIE Optical and IR Telescope Instrumentation and Detectors* Vol. 4008 (eds Iye, M. & Moorwood, A. F.) 623–631 (SPIE, 2000).
58. Castro-Tirado, A. GRB 190114C: refined redshift by the 10.4m GTC. *GCN Circulars* 23708 https://gcn.gsfc.nasa.gov/gcn3/23708.gcn3 (2019).

59. de Ugarte Postigo, A. et al. The distribution of equivalent widths in long GRB afterglow spectra. *Astron. Astrophys.* **548**, A11 (2012).
60. Steele, I. A. et al. The Liverpool Telescope: performance and first results. In *Proc. SPIE Ground-based Telescopes* Vol. 5489 (ed. Oschmann, J. M. Jr) 679–692 (SPIE, 2004).
61. Chambers, K. C. et al. The Pan-STARRS1 surveys. Preprint at https://arxiv.org/abs/1612.05560 (2016).
62. Tarenghi, M. & Wilson, R. N. The ESO NTT (New Technology Telescope): the first active optics telescope. In *Proc. SPIE Active Telescope Systems* Vol. 1114 (ed. Roddier, F. J.) 302–313 (SPIE, 1989).
63. Smartt, S. J. et al. PESSTO: survey description and products from the first data release by the Public ESO Spectroscopic Survey of Transient Objects. *Astron. Astrophys.* **579**, A40 (2015).
64. Covino, S. et al. REM: a fully robotic telescope for GRB observations. In *Proc. SPIE Ground-based Instrumentation for Astronomy* Vol. 5492 (eds Moorwood, A. F. M. & Iye, M.) 1613–1622 (SPIE, 2004).
65. Roming, P. W. A. et al. The Swift ultra-violet/optical telescope. *Space Sci. Rev.* **120**, 95–142 (2005).
66. Siegel, M. H. & Gropp, J. D. GRB 190114C: Swift/UVOT detection. *GCN Circulars* 23725 https://gcn.gsfc.nasa.gov/gcn3/23725.gcn3 (2019).
67. Poole, T. S. et al. Photometric calibration of the Swift ultraviolet/optical telescope. *Mon. Not. R. Astron. Soc.* **383**, 627–645 (2008).
68. Breeveld, A. A. et al. An updated ultraviolet calibration for the Swift/UVOT. In *American Institute of Physics Conference Series* Vol. 1358, 373–376 (AIP, 2011).
69. Kuin, N. P. M. et al. Calibration of the Swift-UVOT ultraviolet and visible grisms. *Mon. Not. R. Astron. Soc.* **449**, 2514–2538 (2015).
70. Arnouts, S. et al. Measuring and modelling the redshift evolution of clustering: the Hubble Deep Field North. *Mon. Not. R. Astron. Soc.* **310**, 540–556 (1999).
71. Ilbert, O. et al. Accurate photometric redshifts for the CFHT legacy survey calibrated using the VIMOS VLT deep survey. *Astron. Astrophys.* **457**, 841–856 (2006).
72. Covino, S. et al. Dust extinctions for an unbiased sample of gamma-ray burst afterglows. *Mon. Not. R. Astron. Soc.* **432**, 1231–1244 (2013).
73. Schlafly, E. F. & Finkbeiner, D. P. Measuring reddening with Sloan Digital Sky Survey stellar spectra and recalibrating SFD. *Astrophys. J.* **737**, 103 (2011).
74. McMullin, J. P., Waters, B., Schiebel, D., Young, W. & Golap, K. CASA architecture and applications. In *Astronomical Data Analysis Software and Systems XVI*, Vol. 376 (eds Shaw, R. A. et al.) 127 (ASP, 2007).
75. Wilson, W. E. et al. The Australia Telescope Compact Array broad-band backend: description and first results. *Mon. Not. R. Astron. Soc.* **416**, 832–856 (2011).
76. Sault, R. J., Teuben, P. J. & Wright, M. C. H. A retrospective view of MIRIAD. In *Astronomical Data Analysis Software and Systems IV* Vol. 77 (eds Shaw, R. A. et al.) 433 (ASP, 1995).
77. Swarup, G. et al. The Giant Metre-wave Radio Telescope. *Current Science* **60**, 95–105 (1991).
78. Cherukuri, S. V. et al. GRB 190114C: GMRT detection at 1.26GHz. *GCN Circulars* 23762 https://gcn.gsfc.nasa.gov/gcn3/23762.gcn3 (2019).
79. Tremou, L. et al. GRB 190114C: MeerKAT radio observation. *GCN Circulars* 23760 https://gcn.gsfc.nasa.gov/gcn3/23760.gcn3 (2019).
80. Camilo, F. et al. Revival of the magnetar PSR J1622-4950: observations with MeerKAT, Parkes, XMM-Newton, Swift, Chandra, and NuSTAR. *Astrophys. J.* **856**, 180 (2018).
81. Jonas, J. L. & The MeerKAT Team. The MeerKAT Radio Telescope. In *Proc. of MeerKAT Science: On the Pathway to the SKA* 001 (2016).
82. Fender, R. et al. ThunderKAT: the MeerKAT large survey project for image-plane radio transients. Preprint at https://arxiv.org/abs/1711.04132 (2017).
83. Mohan, N. & Rafferty, D. PyBDSF: Python Blob Detection and Source Finder https://www.astron.nl/citt/pybdsf/ (2015)
84. Holland, W. S. et al. SCUBA-2: the 10 000 pixel bolometer camera on the James Clerk Maxwell Telescope. *Mon. Not. R. Astron. Soc.* **430**, 2513–2533 (2013).
85. Bošnjak, Ž., Daigne, F. & Dubus, G. Prompt high-energy emission from gamma-ray bursts in the internal shock model. *Astron. Astrophys.* **498**, 677–703 (2009).
86. Panaitescu, A. & Kumar, P. Analytic light curves of gamma-ray burst afterglows: homogeneous versus wind external media. *Astrophys. J.* **543**, 66–76 (2000).
87. Derishev, E. & Piran, T. The physical conditions of the afterglow implied by MAGIC's sub-TeV observations of GRB 190114C. *Astrophys. J. Lett.* **880**, 27 (2019).
88. Mastichiadis, A. & Kirk, J. G. Self-consistent particle acceleration in active galactic nuclei. *Astron. Astrophys.* **295**, 613 (1995).
89. Vurm, I. & Poutanen, J. Time-dependent modeling of radiative processes in hot magnetized plasmas. *Astrophys. J.* **698**, 293–316 (2009).
90. Petropoulou, M. & Mastichiadis, A. On the multiwavelength emission from gamma ray burst afterglows. *Astron. Astrophys.* **507**, 599–610 (2009).
91. Pennanen, T., Vurm, I. & Poutanen, J. Simulations of gamma-ray burst afterglows with a relativistic kinetic code. *Astron. Astrophys.* **564**, A77 (2014).

Acknowledgements We thank the Instituto de Astrofísica de Canarias for the excellent working conditions at the Observatorio del Roque de los Muchachos in La Palma. We acknowledge financial support by the German BMBF and MPG, the Italian INFN and INAF, the Swiss National Fund SNF, the ERDF under the Spanish MINECO (FPA2017-87859-P, FPA2017-85668-P, FPA2017-82729-C6-2-R, FPA2017-82729-C6-6-R, FPA2017-82729-C6-5-R, AYA2015-71042-P, AYA2016-76012-C3-1-P, ESP2017-87055-C2-2-P, FPA201790566REDC), the Indian Department of Atomic Energy, the Japanese JSPS and MEXT, the Bulgarian Ministry of Education and Science, National RI Roadmap Project DO1-153/28.08.2018 and the Academy of Finland grant number 320045. This work was also supported by the Spanish Centro de Excelencia 'Severo Ochoa' through grants SEV-2016-0588 and SEV-2015-0548 and Unidad de Excelencia 'María de Maeztu' MDM-2014-0369, by the Croatian Science Foundation (HrZZ) Project IP-2016-06-9782 and the University of Rijeka Project 13.12.1.3.02, by the DFG Collaborative Research Centers SFB823/C4 and SFB876/C3, the Polish National Research Centre grant UMO-2016/22/M/ST9/00382 and by the Brazilian MCTIC, CNPq and FAPERJ. L. Nava acknowledges funding from the European Union's Horizon 2020 Research and Innovation programme under the Marie Skłodowska-Curie grant agreement number 664931. E. Moretti acknowledges funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement number 665919. This study used the following ALMA data: ADS/JAO.ALMA#2018.A.00020.T, ADS/JAO.ALMA#2018.1.01410.T. ALMA is a partnership of ESO (representing its member states), NSF (USA) and NINS (Japan), together with NRC (Canada), MOST and ASIAA (Taiwan), and KASI (Republic of Korea), in cooperation with the Republic of Chile. The Joint ALMA Observatory is operated by ESO, AUI/NRAO and NAOJ. C.C.T., A.d.U.P. and D.A.K. acknowledge support from the Spanish research project AYA2017-89384-P. C.C.T and A.d.U.P. acknowledge support from funding associated with Ramón y Cajal fellowships (RyC-2012-09984 and RyC-2012-09975). D.A.K. acknowledges support from funding associated with Juan de la Cierva Incorporación fellowships (IJCI-2015-26153). The JCMT is operated by the East Asian Observatory on behalf of The National Astronomical Observatory of Japan, Academia Sinica Institute of Astronomy and Astrophysics, the Korea Astronomy and Space Science Institute, and Center for Astronomical Mega-Science (as well as the National Key R&D Program of China via grant number 2017YFA0402700). Additional funding support is provided by the Science and Technology Facilities Council of the UK and participating universities in the UK and Canada. The JCMT data reported here were obtained under project M18BP040 (principal investigator D.A.P.). We thank M. Rawlings, K. Silva, S. Urquart and the JCMT staff for support for these observations. The Liverpool Telescope, located on the island of La Palma, in the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofisica de Canarias, is operated by Liverpool John Moores University with financial support from the UK Science and Technology Facilities Council. The Australia Telescope Compact Array is part of the Australia Telescope National Facility, which is funded by the Australian Government for operation as a National Facility managed by CSIRO. G.E.A. is the recipient of an Australian Research Council Discovery Early Career Researcher Award (project number DE180100346) and J.C.A.M.-J. is the recipient of an Australian Research Council Future Fellowship (project number FT140101082) funded by the Australian Government. Support for the German contribution to GBM was provided by the Bundesministerium für Bildung und Forschung (BMBF) via the Deutsches Zentrum für Luft und Raumfahrt (DLR) under grant number 50 QV 0301. The University of Alabama in Huntsville (UAH) coauthors acknowledge NASA funding from cooperative agreement NNM11AA01A. C.A.W.-H. and C.M.H. acknowledge NASA funding through the Fermi-GBM project. The Fermi LAT Collaboration acknowledges support from a number of agencies and institutes that have supported both the development and the operation of the LAT, as well as scientific data analysis. These include the National Aeronautics and Space Administration and the Department of Energy (DOE) in the USA; the Commissariat à l'Energie Atomique and the Centre National de la Recherche Scientifique/Institut National de Physique Nucléaire et de Physique des Particules in France; the Agenzia Spaziale Italiana and the Istituto Nazionale di Fisica Nucleare in Italy; the Ministry of Education, Culture, Sports, Science and Technology (MEXT), High Energy Accelerator Research Organization (KEK) and Japan Aerospace Exploration Agency (JAXA) in Japan; and the K. A. Wallenberg Foundation, the Swedish Research Council and the Swedish National Space Board in Sweden. We acknowledge additional support for science analysis during the operations phase from the Istituto Nazionale di Astrofisica in Italy and the Centre National d'Études Spatiales in France. This work was performed in part under DOE contract DE-AC02-76SF00515. Part of the funding for GROND (both hardware and personnel) was granted from the Leibniz-Prize to G. Hasinger (DFG grant HA 1850/28-1). Swift data were retrieved from the Swift archive at HEASARC/NASA-GSFC and from the UK Swift Science Data Centre. Support for Swift in the UK is provided by the UK Space Agency. This work is based on observations obtained with XMM-Newton, an ESA science mission with instruments and contributions directly funded by ESA Member States and NASA. This work is partially based on observations collected at the European Organisation for Astronomical Research in the Southern Hemisphere under ESO programme 199.D-0143. The work is partly based on observations made with the GTC, installed in the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofísica de Canarias, in the island of La Palma. This work is partially based on observations made with the NOT (programme 58-502), operated by the Nordic Optical Telescope Scientific Association at the Observatorio del Roque de los Muchachos, La Palma, Spain, of the Instituto de Astrofísica de Canarias. This work is partially based on observations collected at the European Organisation for Astronomical Research in the Southern Hemisphere under ESO programme 102.D-0662. This work is partially based on observations collected through the ESO programme 199.D-0143 ePESSTO. M. Gromadzki is supported by the Polish NCN MAESTRO grant 2014/14/A/ST9/00121. M.N. is supported by a Royal Astronomical Society Research Fellowship M.G.B., S. Campana, A. Melandri and P.D'A. acknowledge ASI grant I/004/11/3. S. Campana acknowledges support from agreement ASI-INAF number 2017-14-H.0. S.J.S. acknowledges funding from STFC grant ST/P000312/1. N.P.M.K. acknowledges support by the UK Space Agency under grant ST/P002323/1 and the UK Science and Technology Facilities Council under grant ST/N00811/1. L.P. and S. Lotti acknowledge partial support from agreement ASI-INAF number 2017-14-H.0. A.F.V. acknowledges RFBR 18-29-21030 for support. A.J.C.-T. acknowledges support from the Junta de Andalucía (Project P07-TIC-03094) and from the Spanish Ministry Projects AYA2012-39727-C03-01 and 2015-71718R. K. Misra acknowledges support from the Department of Science and Technology (DST), Government of India and the Indo-US Science and Technology Forum (IUSSTF) for the WISTEMM fellowship and Departnment of Physics, UC Davis, where a part of this work was carried out. S.B.P. and K. Misra acknowledge BRICS (Brazil, Russia, India, China and South Africa) grant DST/IMRCD/BRICS/Pilotcall/ProFCheap/2017(G) for this work. M.J.M. acknowledges the support of the National Science Centre, Poland, through grant 2018/30/E/ST9/00208. V.J. and L.R. acknowledge support from grant EMR/2016/007127 from the Department of Science and Technology, India. K. Maguire acknowledges support from H2020 through an ERC starting grant (758638). L.I. acknowledges M. Della Valle for support in the operation of the telescope.

Author contributions The MAGIC telescope system was designed and constructed by the MAGIC Collaboration. Operation, data processing, calibration, Monte Carlo simulations of the detector and of theoretical models, and data analyses were performed by the members of the MAGIC Collaboration, who also discussed and approved the scientific results. L. Nava coordinated the collection of the data, developed the theoretical interpretation and wrote the main section and the section on afterglow modelling. E. Moretti coordinated the analysis of the MAGIC data, wrote the relevant sections and, together with F. Longo, coordinated the

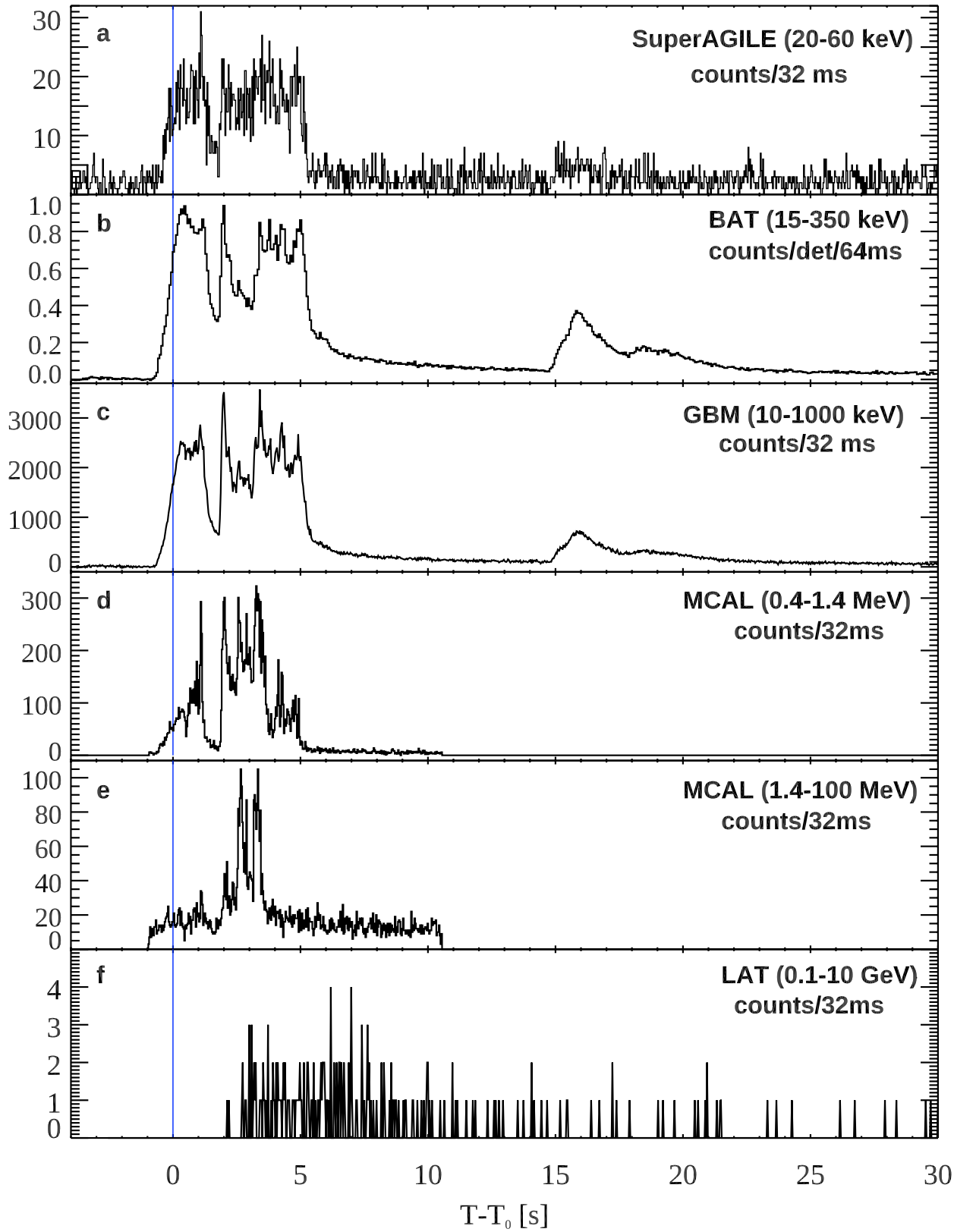
collaboration with the Fermi team. D. Miceli, Y.S. and S.F. performed the analysis of the MAGIC data. S. Covino provided support with the analysis of the optical data and the writing of the corresponding sections. Z.B. performed calculations for the contribution of prompt emission to the teraelectronvolt radiation and wrote the corresponding section. A. Stamerra, D.P. and S.I. contributed to structuring and editing the paper. A. Berti contributed to editing and finalizing the manuscript. R.M. coordinated and supervised the writing of the paper. All MAGIC collaborators contributed to the editing of and provided comments on the final version of the manuscript. S. Campana and M.G.B. extracted the spectra and performed the spectral analysis of the Swift-BAT and Swift-XRT data. N.P.M.K. derived the photometry for the Swift-UVOT event mode data and the UV grism exposure. M.H.S. derived the image-mode Swift-UVOT photometry. A.d.U.P. was principal investigator of ALMA programme 2018.1A.00020.T, triggered these observations and performed photometry. S. Martin reduced the ALMA Band 6 data. C.C.T., S. Schulze, D.A.K. and M. Michałowski participated in the ALMA DDT proposal preparation, observations and scientific analysis of the data. D.A.P. was principal investigator of ALMA programme 2018.1.01410.T and triggered these observations and was principal investigator of the LT and JCMT programmes. A.M.C. analysed the ALMA Band 3 and LT data and wrote the LT text. S. Schulze contributed to the development of the ALMA Band 3 observing programme. I.A.S. triggered the JCMT programme, analysed the data and wrote the associated text. N.R.T. contributed to the development of the JCMT programme. D.A.K. and C.C.T. triggered and coordinated the X-shooter observations. D.A.K. independently checked the optical light curve analysis. K. Misra was the principal investigator of the GMRT programme 35_018. S.V.C. and V.J. analysed the data. L.R. contributed to the observation plan and data analysis. E. Tremou, I.H. and R.D. performed the MeerKAT data analysis. G.E.A., A. Moin, S. Schulze and E. Troja were principal investigators of ATCA programme CX424. G.E.A., M. Wieringa and J. Stevens carried out the observations. G.E.A., G. Bernardi, S.K., M. Marongiu, A. Moin, R.R. and M. Wieringa analysed these data. J.C.A.M.-J. and L.P. participated in the ATCA proposal preparation and the scientific analysis of the data. The ePESSTO project was delivered by the following, who contributed to managing, executing, reducing, analysing ESO/NTT data and provided comments to the manuscript: J.P.A., N.C.S., P.D'A., M. Gromadzki, C.I., E.K., K. Maguire, M.N., F.R. and S.J.S.; A. Melandri and A. Rossi reduced and analysed REM data and provided comments to the manuscript. J. Bolmer was responsible for observing the GRB with GROND and for the data reduction and calibration. J. Bolmer and J. Greiner contributed to the analysis of the data and writing of the text. E. Troja triggered the NuSTAR TOO observations performed under the DDT programme, L.P. requested the XMM-Newton data, obtained under a DDT programme, and carried out the scientific analysis of the XMM-Newton and NuSTAR data. S. Lotti analysed the NuSTAR data and wrote the associated text. A. Tiengo and G. Novara analysed the XMM-Newton data and wrote the associated text. A.J.C.-T. led the observing BOOTES and GTC programmes. A. Castellón, C.J.P.d.P., E.F.-G., I.M.C., S.B.P. and X.Y.L. analysed the BOOTES data, and A.F.V., M.D.C.-G., R.S.-R., Y.-D.H. and V.V.S. analysed the GTC data and interpreted them accordingly. N.R.T. created the X-shooter and AlFOSC figures. J.P.U.F. and J.J. performed the analysis of the X-shooter and AlFOSC spectra. D.X. and P.J. contributed to the NOT programme and triggering. D. Malesani performed photometric analysis of NOT data. E. Peretti contributed to the development of the code for modelling afterglow radiation. L.I. triggered and analysed the OASDG data, and A.D.D. and A.N. performed the observations at the telescope.

Competing interests The authors declare no competing interests.

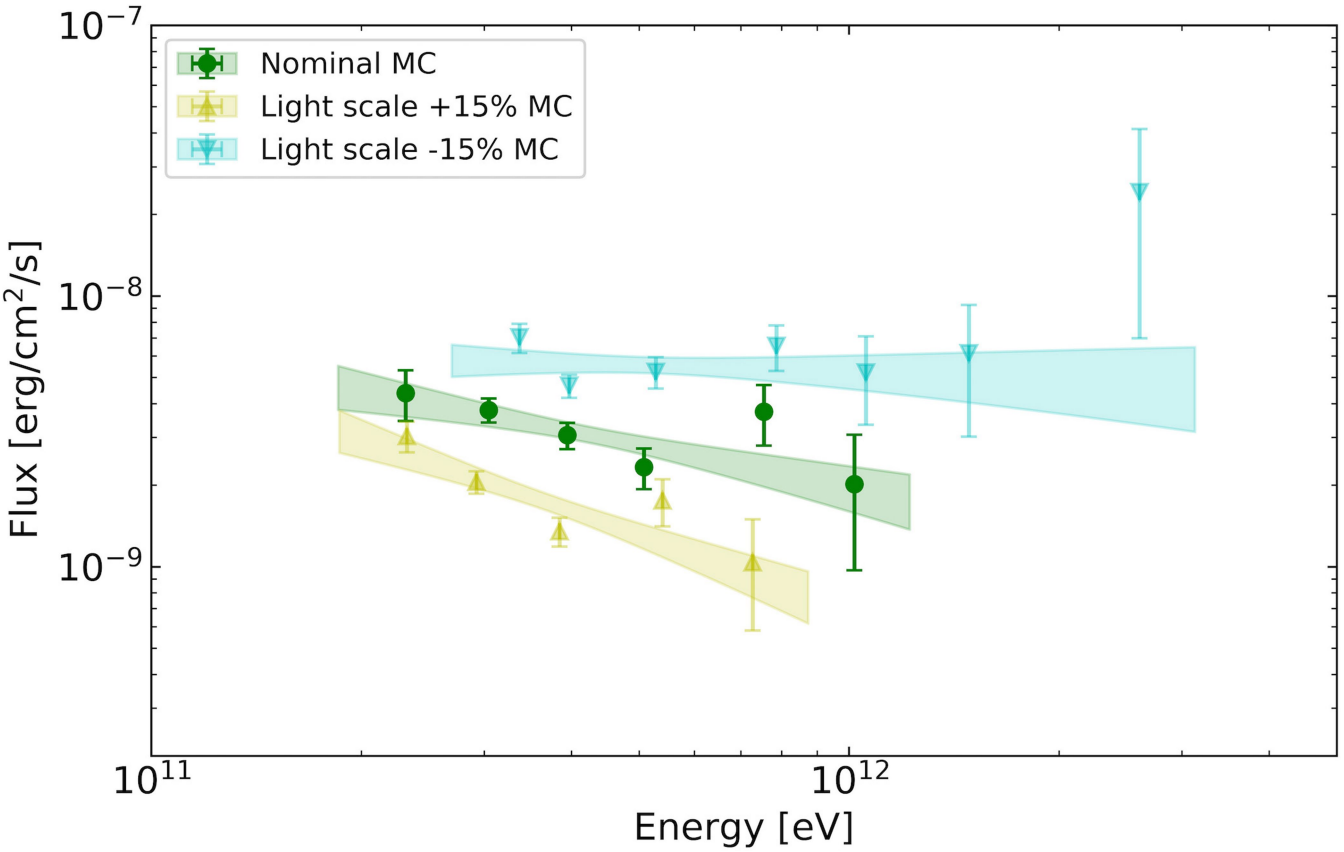
Additional information

Correspondence and requests for materials should be addressed to R.M.

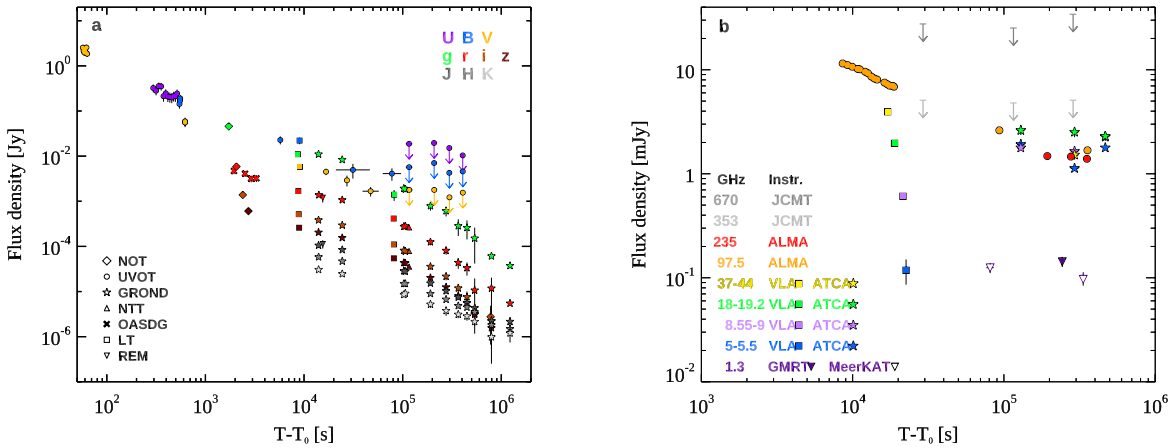
Peer review information *Nature* thanks Xiang-Yu Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



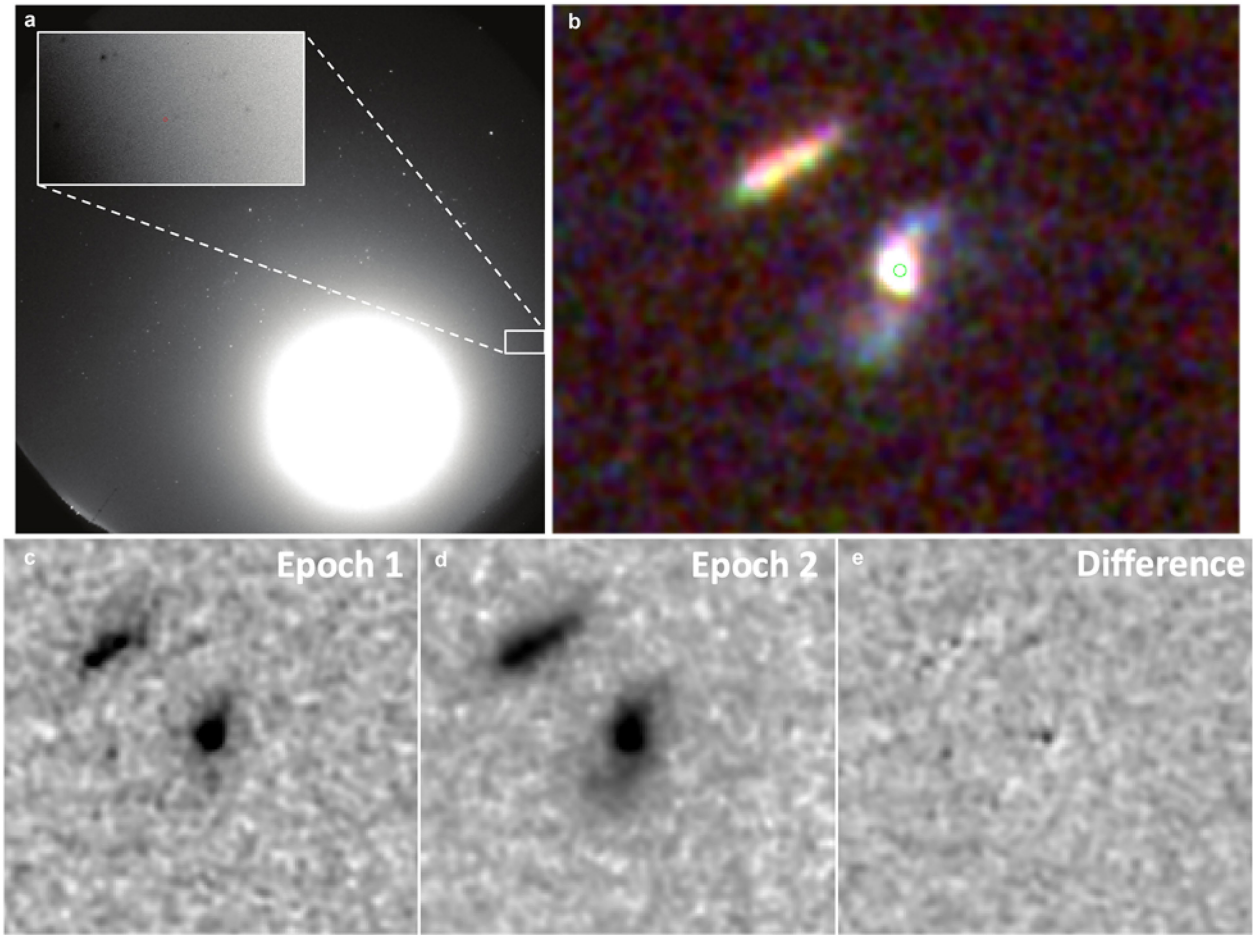
Extended Data Fig. 1 | Prompt-emission light curves for different detectors. **a**–**f**, Light curves for Super-AGILE (**a**; 20–60 keV), Swift-BAT (**b**; 15–150 keV), Fermi-GBM (**c**; 10–1,000 keV), AGILE-MCAL (**d**; 0.4–1.4 MeV), AGILE-MCAL (**e**; 1.4–100 MeV) and Fermi-LAT (**f**; 0.1–10 GeV). The light curve of AGILE-MCAL is split into two bands to show the energy dependence of the first peak. Error bars show 1σ statistical errors.



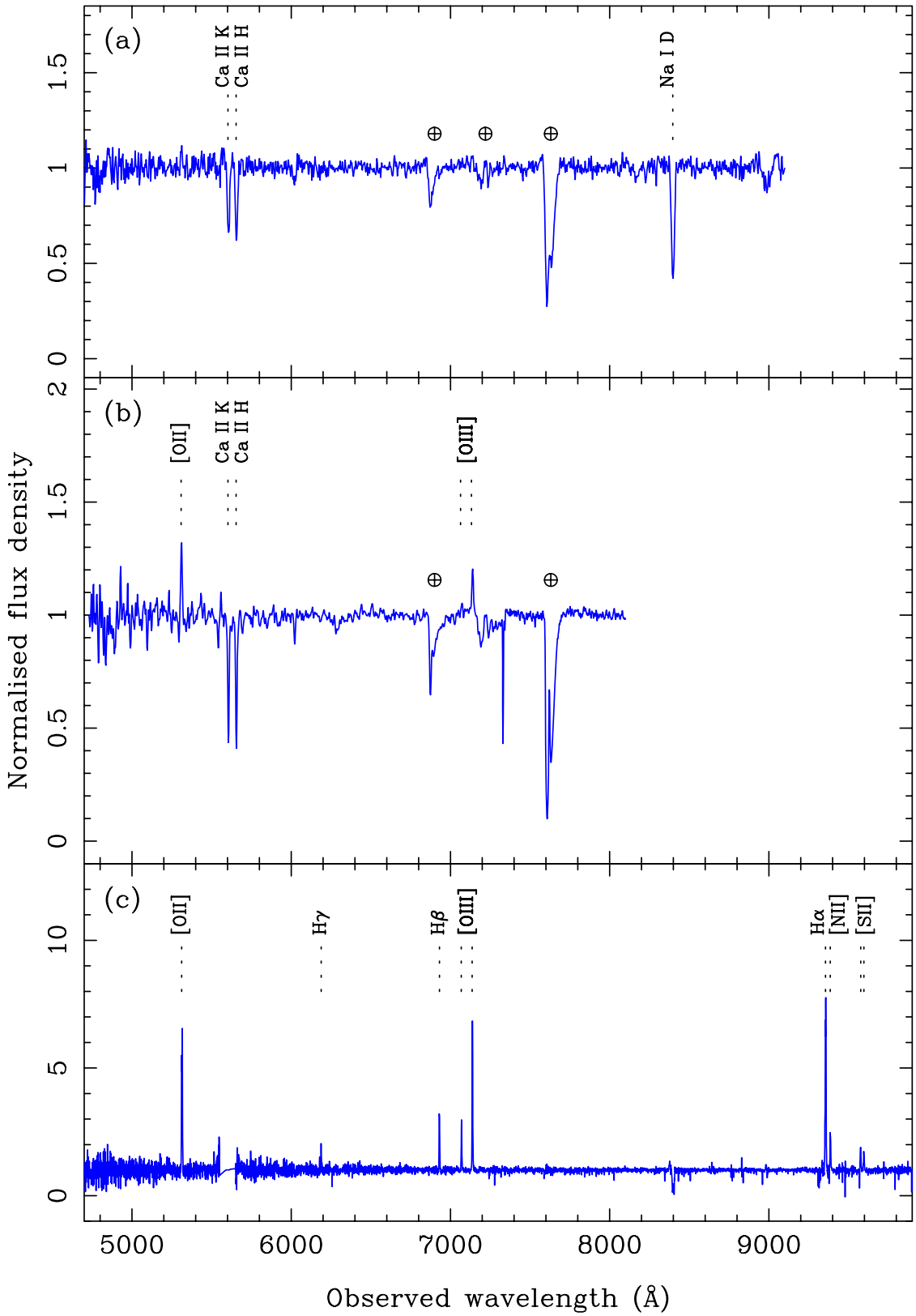
Extended Data Fig. 2 | MAGIC time-integrated SEDs in the time interval 62–2,400 s after T_0. The green (yellow, blue) points and band show the results of the Monte Carlo (MC) simulations for the nominal and the varied light scale cases (+15%, −15%), which define the limits of the systematic uncertainties. The contour regions are drawn from the 1σ error of their best-fit power-law functions. The vertical bars of the data points show the 1σ errors on the flux.



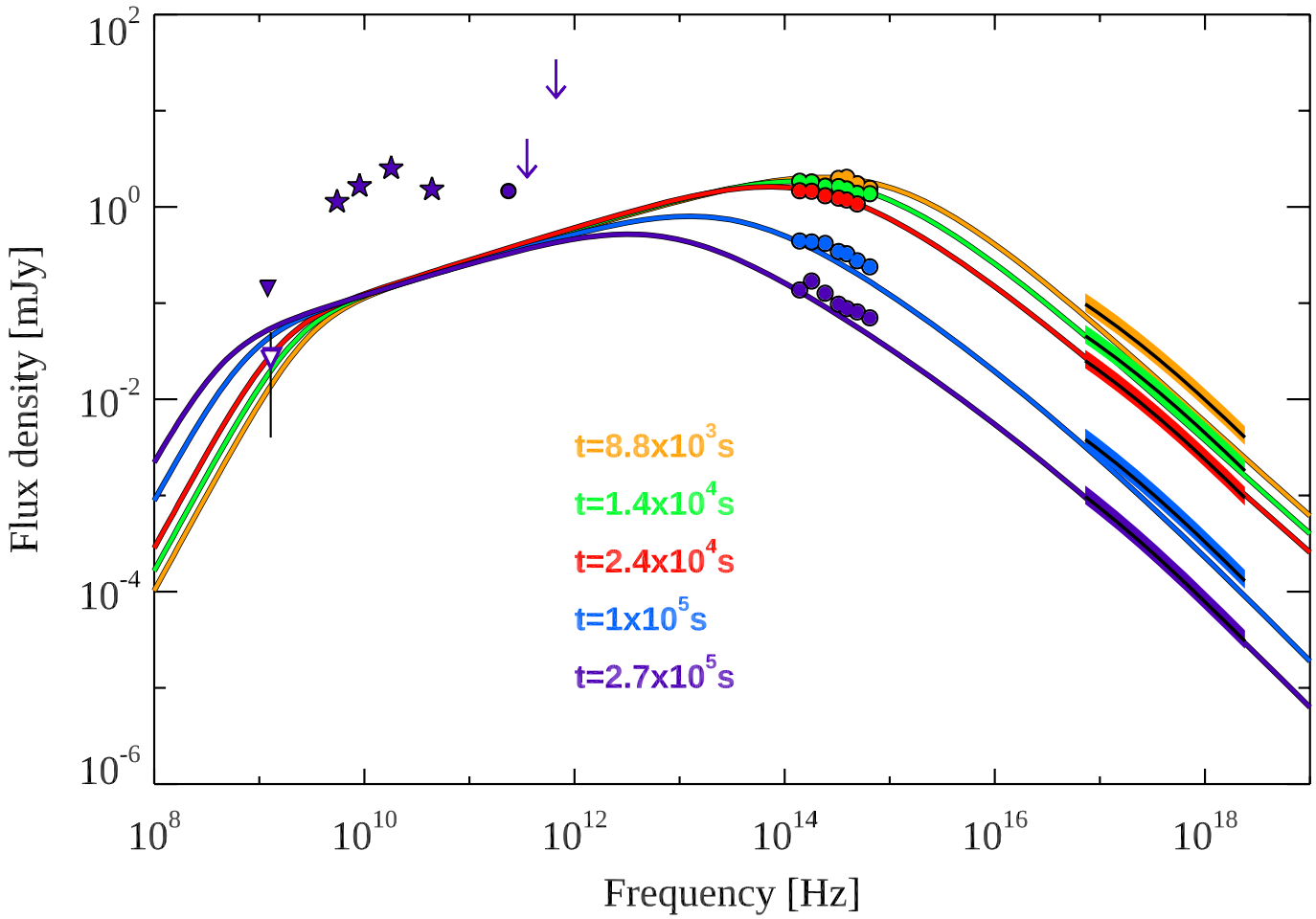
Extended Data Fig. 3 | Afterglow light curves of GRB 190114C. Flux density at different frequencies as a function of the time since the initial burst, $T - T_0$. **a**, Observation in the NIR, optical and UV bands. The flux has been corrected for extinction in the host and in our Galaxy. The contribution of the host galaxy and its companion has been subtracted. Fluxes have been rescaled (except for the r-band filter). **b**, Radio and submillimetre observations from 1.3 GHz to 670 GHz. 'Instr.', instrument.



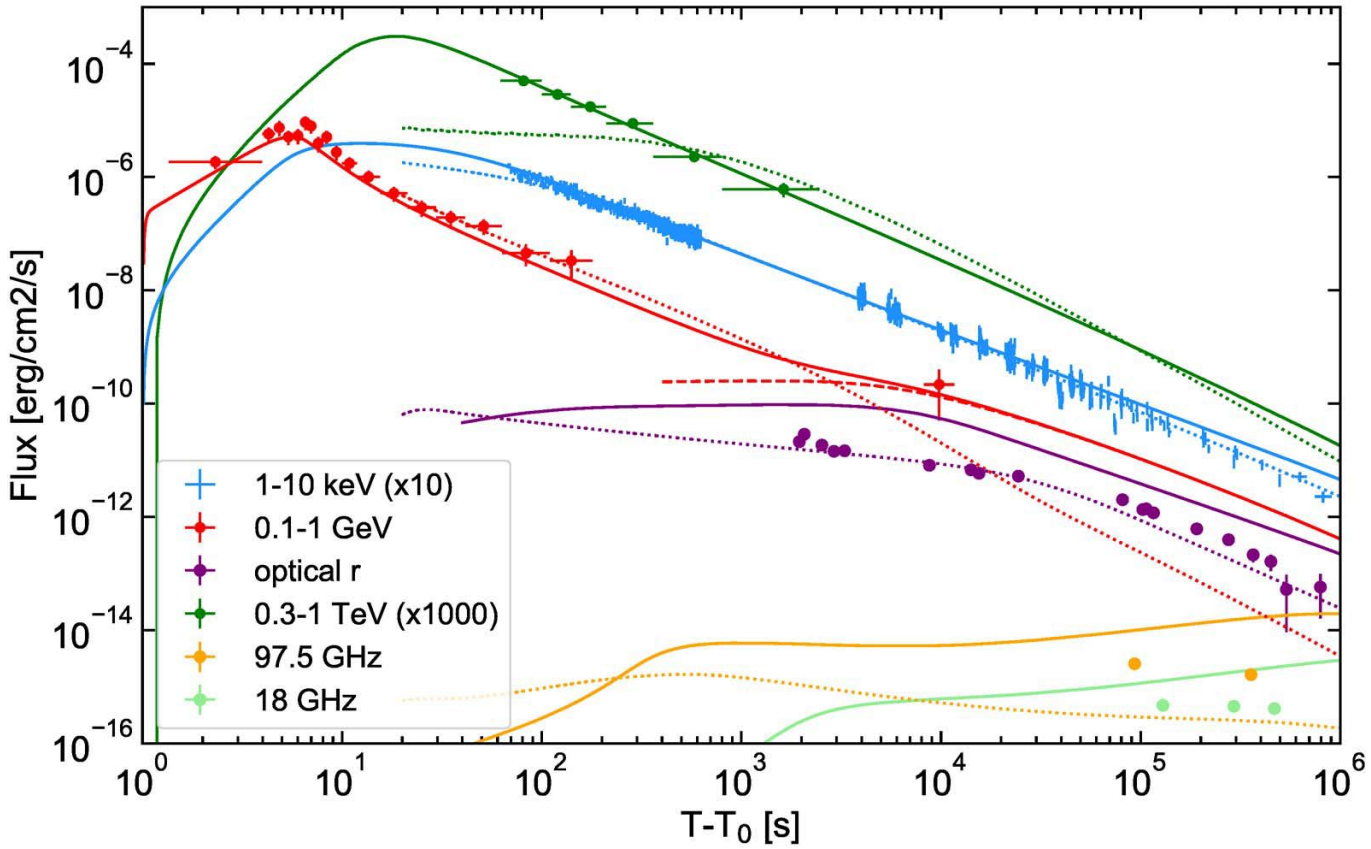
Extended Data Fig. 4 | Images of the localization region of GRB 190114C. **a**, All-sky image captured with the CASANDRA-1 camera at the BOOTES-1 station. The image (30 s exposure, unfiltered) was taken at T_0 + 14.8 s, and was severely affected by the moon. At the GRB190114C location (red dot) no prompt optical emission is detected. Inset, magnification (inverted colours) containing a 10′-diametercircle centred on the optical position. **b**, Three-colour image of the host of GRB 190114C, obtained with the HST. The host galaxy is a spiral galaxy, and the green circle indicates the location of the transient close to its host nucleus. The image is 8″ across; north is up and east is to the left. **c**–**e**, Images of the GRB 190114C field taken with the HST, obtained with the F850LP filter (covering roughly the region from 800 to 1,100 nm). Two epochs, 11 February and 12 March 2019, are shown (images are 4″ across); the right-most image is the result of the difference image. A faint transient is visible close to the nucleus of the galaxy, and we identify this as the late-time afterglow of the burst.



Extended Data Fig. 5 | Optical–NIR spectra of GRB 190114C. a, NOT/AlFOSC spectrum obtained at mid-time (i.e., the epoch corresponding to a half of the exposure length) 1 h post-burst. The continuum is afterglow-dominated at this time, and shows strong absorption features of Ca II and Na I (in addition to telluric absorption). **b**, Normalized GTC (+OSIRIS) spectrum obtained on 14 January 2019, 23:32:03 UT with the R1000B and R2500I grisms. The emission lines of the underlying host galaxy are noticeable, besides the Ca II absorption lines in the afterglow spectrum. **c**, Visible-light region of the VLT–X-shooter spectrum obtained approximately 3.2 d post-burst, showing strong emission lines from the star-forming host galaxy.



Extended Data Fig. 6 | SEDs from radio frequencies to X-rays at different epochs. The synchrotron frequency ν_m crosses the optical band, moving from higher to lower frequencies. The break between 10^8 and 10^{10} Hz is caused by the self-absorption synchrotron frequency, ν_{sa}. Optical (X-ray) data have been corrected for extinction (absorption). The data points are taken from the following telescopes (from lower to higher frequencies): filled and empty triangle symbols, GMRT and MeerKAT; stars, ATCA; violet filled circle, ALMA, down arrows, JCMT 1σ upper limits; filled circles, LT (yellow) and GROND (all the other colours). Error bars for all data points define the 1σ error. Coloured stripes show the best fit of the XRT data extrapolated to the time of each SED. Their vertical width is obtained from the error (90% confidence level) on the best-fit normalization. Solid lines show the model SEDs for the case $s = 2$.



Extended Data Fig. 7 | Modelling of broadband light curves. Modelling results of forward shock emission are compared to observations at different frequencies (see key). The model shown with solid and dashed lines is optimized to describe the high-energy radiation (teraelectronvolt, gigaelectronvolt and X-ray) and has been obtained with the following parameters: $s = 0$, $\varepsilon_e = 0.07$, $\varepsilon_B = 8 \times 10^{-5}$, $p = 2.6$, $n_0 = 0.5$ and $E_k = 8 \times 10^{53}$ erg. Solid lines show the total flux (synchrotron and SSC) and the dashed line refers to the SSC contribution only. Dotted curves correspond to a better modelling of observations at lower frequencies, but fail to explain the behaviour of the teraelectronvolt light curve; they are obtained with the following model parameters: $s = 2$, $\varepsilon_e = 0.6$, $\varepsilon_B = 10^{-4}$, $p = 2.4$, $A. = 0.1$ and $E_k = 4 \times 10^{53}$ erg. Vertical bars on the data points show the 1σ errors on the flux, and horizontal bars represent the duration of the observation.

Extended Data Table 1 | MAGIC spectral-fit parameters for GRB 190114C

Time bin [seconds after T_0]	Normalisation [$\mathrm{TeV}^{-1}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$]	Photon index	Pivot energy [GeV]
62 – 90	$1.95^{+0.21}_{-0.20} \cdot 10^{-7}$	$-2.17^{+0.34}_{-0.36}$	395.5
68 – 180	$1.10^{+0.09}_{-0.08} \cdot 10^{-7}$	$-2.27^{+0.24}_{-0.25}$	404.7
180 – 625	$2.26^{+0.21}_{-0.20} \cdot 10^{-8}$	$-2.56^{+0.27}_{-0.29}$	395.5
68 – 110	$1.74^{+0.16}_{-0.15} \cdot 10^{-7}$	$-2.16^{+0.29}_{-0.31}$	386.5
110 – 180	$8.59^{+0.95}_{-0.91} \cdot 10^{-8}$	$-2.51^{+0.37}_{-0.41}$	395.5
180 – 360	$3.50^{+0.38}_{-0.36} \cdot 10^{-8}$	$-2.36^{+0.34}_{-0.37}$	395.5
360 – 625	$1.65^{+0.23}_{-0.23} \cdot 10^{-8}$	$-3.16^{+0.48}_{-0.54}$	369.1
625 – 2400	$3.52^{+0.47}_{-0.47} \cdot 10^{-9}$	$-2.80^{+0.48}_{-0.54}$	369.1
62 – 2400 (Nominal MC)	$1.07^{+0.08}_{-0.07} \cdot 10^{-8}$	$-2.51^{+0.20}_{-0.21}$	423.8
62 – 2400 (Light scale +15% MC)	$7.95^{+0.58}_{-0.56} \cdot 10^{-9}$	$-2.91^{+0.23}_{-0.25}$	369.1
62 – 2400 (Light scale -15% MC)	$1.34^{+0.09}_{-0.09} \cdot 10^{-8}$	$-2.07^{+0.18}_{-0.19}$	509.5

For each time bin, the table shows the start and end time of the bin, the normalization factor of the EBL-corrected differential flux at the pivot energy with statistical errors, photon indices with statistical errors, and the pivot energy of the fit (fixed).

Extended Data Table 2 | GROND photometry

T_{GROND}	AB magnitude						
(s)	g'	r'	i'	z'	J	H	K_s
14029.94 ± 335.28	19.21 ± 0.03	18.46 ± 0.03	17.78 ± 0.03	17.33 ± 0.03	16.78 ± 0.05	16.30 ± 0.05	16.03 ± 0.07
24402.00 ± 345.66	19.50 ± 0.04	18.72 ± 0.03	18.05 ± 0.03	17.61 ± 0.03	17.02 ± 0.05	16.53 ± 0.05	16.26 ± 0.08
102697.17 ± 524.01	20.83 ± 0.06	20.00 ± 0.04	19.30 ± 0.04	18.87 ± 0.03	18.15 ± 0.05	17.75 ± 0.06	17.40 ± 0.09
106405.63 ± 519.87	20.86 ± 0.05	19.98 ± 0.03	19.34 ± 0.03	18.88 ± 0.03	18.17 ± 0.06	17.75 ± 0.06	17.34 ± 0.09
191466.77 ± 751.37	21.43 ± 0.07	20.61 ± 0.03	19.97 ± 0.03	19.52 ± 0.03	18.77 ± 0.06	18.28 ± 0.06	17.92 ± 0.14
275594.19 ± 747.59	21.57 ± 0.07	20.88 ± 0.04	20.31 ± 0.04	19.87 ± 0.04	19.14 ± 0.07	18.57 ± 0.06	18.26 ± 0.21
366390.74 ± 1105.79	21.87 ± 0.07	21.17 ± 0.04	20.62 ± 0.03	20.15 ± 0.03	19.43 ± 0.06	18.89 ± 0.06	18.46 ± 0.15
448791.55 ± 1201.33	21.90 ± 0.08	21.27 ± 0.04	20.79 ± 0.04	20.33 ± 0.03	19.66 ± 0.07	18.97 ± 0.07	18.55 ± 0.18
537481.41 ± 1132.16	22.02 ± 0.09	21.52 ± 0.05	21.00 ± 0.04	20.55 ± 0.03	19.87 ± 0.07	19.20 ± 0.07	18.83 ± 0.17
794992.63 ± 1200.69	22.14 ± 0.04	21.51 ± 0.03	21.05 ± 0.04	20.71 ± 0.05	20.31 ± 0.13	19.79 ± 0.14	19.59 ± 0.41
1226716.84 ± 1050.15	22.17 ± 0.04	21.59 ± 0.04	21.26 ± 0.04	20.97 ± 0.04	20.34 ± 0.12	19.95 ± 0.11	19.40 ± 0.34

Time T_{GROND} after the BAT trigger. The AB magnitudes are not corrected for Galactic foreground reddening.

Extended Data Table 3 | LT, NOT and UVOT observations

UTC	Filter	Exposure (s)	Magnitude
		LT/IO:O	
2019-01-14.975	*g*	45	19.08±0.06
2019-01-14.976	*r*	45	18.22±0.02
2019-01-14.977	*i*	45	17.49±0.02
2019-01-14.978	*z*	45	17.12±0.02
2019-01-14.979	*B*	45	19.55±0.15
2019-01-14.980	*V*	45	18.81±0.08
2019-01-15.814	*r*	60	19.61±0.05
2019-01-15.818	*z*	60	18.70±0.06
2019-01-15.820	*i*	60	19.04±0.04
2019-01-15.823	*g*	60	20.96±0.17
		NOT/AlFOSC	
2019-01-14.89127	*g*	1 × 300	17.72±0.03
2019-01-14.89512	*r*	1 × 300	16.93±0.02
2019-01-14.89899	*i*	1 × 300	16.42 ±0.04
2019-01-14.90286	*z*	1 × 300	16.17 ±0.04
2019-01-23.8896	*i*	6 × 300	21.02±0.05

UVOT							
T_{start}	T_{stop}	Filter	Magnitude	T_{start}	T_{stop}	Filter	Magnitude
56.63	57.63	*V*	12.17±0.14	130958	142524	*UVM2*	20.37
57.63	58.63	*V*	12.34±0.14	217406	222752	*UVM2*	20.48
58.63	59.63	*V*	12.44±0.13	107573	125233	*U*	20.29
59.63	60.63	*V*	12.29±0.14	205500	210750	*U*	20.25
60.63	61.63	*V*	12.44±0.14	291188	302718	*U*	20.49
61.63	62.63	*V*	12.16±0.13	400429	412385	*U*	20.82
62.63	63.63	*V*	12.51±0.13	616	627	*V*	16.25±0.20
615.95	625.95	*V*	16.32±0.20	16295	17136	*V*	19.03±0.14
73.34	83.34	*white*	13.86	26775	27682	*V*	19.50±0.27
83.34	93.34	*white*	14.10±0.06	39149	57221	*V*	20.09±0.23
93.34	103.34	*white*	14.19±0.06	108064	125736	*V*	20.02
103.34	113.34	*white*	14.36±0.06	206689	211356	*V*	20.02
113.34	123.34	*white*	14.64±0.06	292383	303996	*V*	20.42
123.34	133.34	*white*	14.65±0.06	401305	413316	*V*	20.17
133.34	143.34	*white*	14.91±0.06	4044	51522	*UVW1*	21.17
143.34	153.34	*white*	14.99±0.06	131216	142656	*UVW1*	20.47
153.34	163.34	*white*	15.05±0.06	217984	223056	*UVW1*	20.57
163.34	173.34	*white*	15.32±0.06	592	612	*UVW2*	17.65
173.34	183.34	*white*	15.38±0.06	6056	56384	*UVW2*	21.30
183.34	193.34	*white*	15.38±0.06	130699	142346	*UVW2*	20.52
193.34	203.34	*white*	15.59±0.06	216828	222404	*UVW2*	20.55
562.0	572.0	*white*	16.96±0.10	566	586	*white*	16.90±0.07
572.0	582.0	*white*	16.90±0.10	607389	613956	*white*	22.16
535.5	555.5	*B*	17.56±0.21	624452	682416	*white*	21.99±0.18
545.5	565.5	*B*	17.25±0.18	745033	769296	*white*	21.64±0.16
285.9	305.9	*U*	17.35±0.19	818840	837216	*white*	22.50
305.9	325.9	*U*	17.50±0.20	893522	907116	*white*	22.57
325.9	345.9	*U*	17.24±0.18	991065	1004196	*white*	22.49±0.35
345.9	365.9	*U*	17.26±0.18	1077542	1094616	*white*	22.41±0.31
365.9	385.9	*U*	17.80±0.24	1140343	1170336	*white*	22.50
385.9	405.9	*U*	17.64±0.22	1220661	1274376	*white*	22.36±0.29
405.9	425.9	*U*	17.82±0.24	5851	6050	*white*	19.25±0.09
425.9	445.9	*U*	17.84±0.25	21950	22857	*white*	20.25±0.09
445.9	465.9	*U*	17.87±0.25	1353459	1359284	*white*	21.70
465.9	485.9	*U*	17.79±0.24	1502211	1548336	*white*	21.98±0.24
485.9	505.9	*U*	17.81±0.24	1692292	1703935	*white*	22.07
505.9	525.9	*U*	17.65±0.22	2132978	2146056	*white*	22.58
542	561	*B*	17.38±0.14	2299521	2317956	*white*	22.41±0.31
5646	5845	*B*	19.54±0.19	63686	80942	*white*	21.07±0.24
21038	46521	*B*	21.14±0.35	107900	125591	*white*	21.40±0.28
62774	96486	*B*	21.33±0.29	206292	211137	*white*	21.52
107737	125412	*B*	21.00	291984	303556	*white*	21.48±0.23
205896	210944	*B*	20.78	401012	413029	*white*	21.84
291586	303137	*B*	21.29	491973	505356	*white*	22.21±0.24
400721	412707	*B*	21.22	74	224	*white*	14.90±0.02
3839	50615	*UVM2*	20.88±0.28				

Magnitudes are SDSS 'AB-like' for ugriz and 'Vega-like' for all the other filters, and they are not corrected for Galactic extinction. For the UVOT data, magnitudes without uncertainties are upper limits.

Extended Data Table 4 | Observations of the host galaxy

Filter	Host	Companion	Combined
Sloan *u*	23.54	25.74	23.40
Sloan *g*	22.51	23.81	22.21
Sloan *r*	22.13	22.81	21.66
Sloan *i*	21.70	22.27	21.19
Sloan *z*	21.51	21.74	20.87
2MASS *J*	20.98	21.08	20.28
2MASS *H*	20.68	20.82	20.00
2MASS *Ks*	20.45	20.61	19.77

For each filter, the estimated magnitudes are given for the host galaxy of GRB 190114C, the companion and the combination of the two objects.

Extended Data Table 5 | Observations of GRB 190114C by ATCA and JCMT SCUBA-2

ATCA			
Start Date and Time	End Date and Time	Frequency GHz	Flux mJy
1/16/2019 6:47:00	1/16/2019 10:53:00	5.5	1.92±0.06
		9	1.78±0.06
		18	2.62±0.26
1/18/2019 1:45:00	1/18/2019 11:18:00	5.5	1.13±0.04
		9	1.65±0.05
		18	2.52±0.27
		44	1.52±0.15
1/20/2019 3:38	1/20/2019 10:25:00	5.5	1.78±0.06
		9	2.26±0.07
		18	2.30±0.23

JCMT SCUBA-2						
UT Date	Time since trigger (days)	Time on source (hours)	Typical 225 GHz CSO Opacity	Typical elevation (degrees)	850 μm RMS density (mJy/beam)	450 μm RMS density (mJy/beam)
2019-01-15	0.338	1.03	0.026	39	1.7	9.2
2019-01-16	1.338	1.03	0.024	39	1.6	8.4
2019-01-18	3.318	0.95	0.031	37	1.7	11.4

For the ATCA data, the start and end dates and times (UTC) of the observations, the frequency and the flux (1σ error) are reported. For the JCMT SCUBA-2 data, the CSO 225-GHz opacity measures the zenith atmospheric attenuation.

A very-high-energy component deep in the γ-ray burst afterglow

https://doi.org/10.1038/s41586-019-1743-9

Received: 5 June 2019

Accepted: 30 September 2019

Published online: 20 November 2019

A list of authors and affiliations appears at the end of the paper.

Gamma-ray bursts (GRBs) are brief flashes of γ-rays and are considered to be the most energetic explosive phenomena in the Universe[1]. The emission from GRBs comprises a short (typically tens of seconds) and bright prompt emission, followed by a much longer afterglow phase. During the afterglow phase, the shocked outflow—produced by the interaction between the ejected matter and the circumburst medium—slows down, and a gradual decrease in brightness is observed[2]. GRBs typically emit most of their energy via γ-rays with energies in the kiloelectronvolt-to-megaelectronvolt range, but a few photons with energies of tens of gigaelectronvolts have been detected by space-based instruments[3]. However, the origins of such high-energy (above one gigaelectronvolt) photons and the presence of very-high-energy (more than 100 gigaelectronvolts) emission have remained elusive[4]. Here we report observations of very-high-energy emission in the bright GRB 180720B deep in the GRB afterglow—ten hours after the end of the prompt emission phase, when the X-ray flux had already decayed by four orders of magnitude. Two possible explanations exist for the observed radiation: inverse Compton emission and synchrotron emission of ultrarelativistic electrons. Our observations show that the energy fluxes in the X-ray and γ-ray range and their photon indices remain comparable to each other throughout the afterglow. This discovery places distinct constraints on the GRB environment for both emission mechanisms, with the inverse Compton explanation alleviating the particle energy requirements for the emission observed at late times. The late timing of this detection has consequences for the future observations of GRBs at the highest energies.

On 20 July 2018, GRB 180720B triggered the Fermi Gamma-ray Burst Monitor (GBM) at 14:21:39.65 universal time (UT)[5] (T_0) and the Swift Burst Alert Telescope (BAT) 5 s later[6]. Multi-wavelength follow-up observations were performed up to $T_0 + 3 \times 10^5$ s by the European Southern Observatory's Very Large Telescope, which measured a redshift of $z = 0.653$ (ref. [7]). In the high-energy γ-ray band (100 MeV–100 GeV) this GRB was also detected by the Fermi Large Area Telescope (LAT) between T_0 and $T_0 + 700$ s with a maximum photon energy of 5 GeV at $T_0 + 142.4$ s (ref. [8]). No further high-energy emission was detected in the successive observation windows after 700 s. The prompt emission phase of GRB 180720B is extremely bright, ranking seventh in brightness among the over 2,650 GRBs detected by Fermi-GBM so far (see Methods). With a T_{90} (the time in which 90% of the flux is detected) of 48.9 ± 0.4 s, GRB 180720B is categorised as a long GRB (typically associated with the death of massive stars[9]), with an isotropic energy release of $E^{iso} = (6.0 \pm 0.1) \times 10^{53}$ erg (50–300 keV; 1 erg = 10^{-7} J). Observations of this GRB took place using the Swift X-ray Telescope (XRT), identifying a bright afterglow that remained detectable until almost 30 days after T_0 (refs. [10,11]; Fig. 1). In terms of energy flux of the X-ray afterglow (0.3–10 keV, at $T_0 + 11$ h), this GRB ranks second after the exceptional GRB 130427A[3].

Observations with the High Energy Stereoscopic System (H.E.S.S.) array began at T_0 + 10.1 h and lasted for two hours. The data were analysed using methods optimized for the detection of the lowest-energy events, revealing a new γ-ray source (Fig. 2a) with an excess of 119 γ-ray events and a statistical significance of 5.3σ (5.0σ post-trial; see Methods). The γ-ray excess is well fitted by a point-like source model centred at a right ascension of 00h 02 min 7.6 s and a declination of −02°56′06″ (J2000) with a statistical uncertainty of 1.31′, consistent with the measurements at other wavelengths[6,8,12]. To rule out the association of this source with an unknown steady γ-ray emitter (such as an active galactic nucleus) or persistent systematic effects, the GRB region was re-observed under similar conditions 18 days after these observations. In total, 6.75 h of data were analysed, resulting in a sky map consistent with background events (Fig. 2b).

The flux spectrum detected by H.E.S.S. (100–440 GeV) was fitted with a function of the form $F_{obs}(E) = F_{int}(E) \times e^{-\tau(E,z)}$, where the exponential term accounts for the absorption of photons by the extragalactic background light[13], τ is the optical depth and $F_{int}(E) = F_{0,int}(E/E_{0,int})^{-\gamma_{int}}$ is a power law describing the intrinsic source emission. The analysis resulted in a photon index of $\gamma_{int} = 1.6 \pm 1.2$ (statistical) ± 0.4 (systematic) and a flux normalization of $F_{0,int} = (7.52 \pm 2.03\ (\text{statistical})\ ^{+4.53}_{-3.84}(\text{systematic})) \times 10^{-10}$ TeV^{-1} cm^{-2} s^{-1}, evaluated at an energy of $E_{0,int} = 0.154$ TeV (see Methods).

The very-high-energy (VHE) flux, together with measurements at other wavelengths, is shown in Fig. 1. Apart from the exceptionally high flux level, the light curves show a typical power-law behaviour in the X-ray and optical afterglow with a temporal flux decay of the form $F(t) \propto t^{-\alpha}$ with $\alpha_{XRT} = 1.29 \pm 0.01$ and $\alpha_{optical} = 1.24 \pm 0.02$. The spectrum measured by Fermi-LAT (100 MeV–10 GeV) from $T_0 + 55$ s to $T_0 + 700$ s is well fitted by a power-law model with photon index $\gamma_{LAT} = 2.10 \pm 0.10$. The light curve in the same time window is fitted by a power law with a temporal decay index of $\alpha_{LAT} = 1.83 \pm 0.25$. It is worth noting that

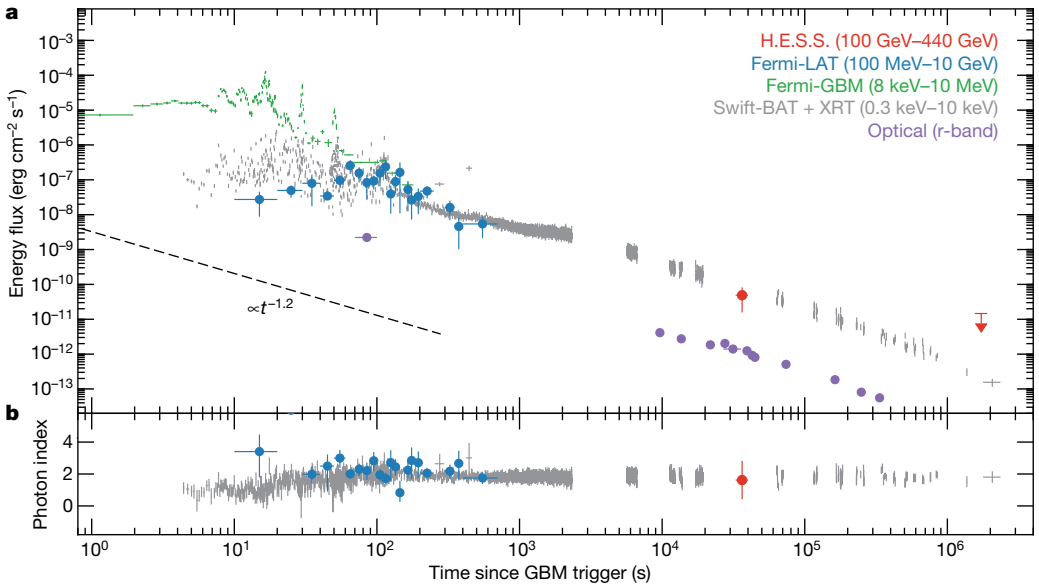


Fig. 1 | Multi-wavelength light curve of GRB 180720B. a, Energy-flux light curve detected by Fermi-GBM (band fit; green), Fermi-LAT (power law; blue), H.E.S.S. (power-law intrinsic; red) and the optical r-band (purple). The Swift-BAT spectra (15 keV–150 keV) are extrapolated to the XRT band (0.3–10 keV) to produce a combined light curve (grey) and an upper limit (95% confidence level) for the second H.E.S.S. observation window (power-law intrinsic, red arrow). The black dashed line indicates a temporal decay with $\alpha = -1.2$. **b**, Photon index of the Fermi-LAT, Swift and H.E.S.S. spectra. Error bars correspond to 1σ.

α_{LAT} is at about 1σ from the mean value of the distribution of the decay indices of long GRBs detected by Fermi-LAT[14] ($\bar{\alpha}_{\mathrm{LAT}} = 0.99 \pm 0.04$, $\sigma_{\bar{\alpha}} = 0.80 + 0.07$) and such deviation could largely depend on the time range in which α_{LAT} is fitted, potentially in agreement with α_{XRT} and $\alpha_{\mathrm{optical}}$.

The detection of VHE γ-ray emission indicates the presence of very energetic particles in the GRB afterglow. This discovery is consistent with efficient γ-ray emission seen in other astrophysical sources with relativistic plasma outflow, for example, pulsar wind nebulae or jets emerging from the nuclei of active galaxies. In the case of a GRB afterglow, the particle acceleration probably occurs at the forward shock[15] (the compression shock wave propagating through the circumburst material), which should be capable of efficient electron and proton acceleration. As proton radiation processes are characterized by long energy-loss timescales relative to the dynamical timescale, the detected γ-ray emission is probably produced by accelerated electrons (see Methods). Therefore, two radiation processes are the most plausible dominant contributions to the VHE spectrum: synchrotron emission of an electron population in the local magnetic field[16] and synchrotron self-Compton (SSC) scattering[17,18]. In the latter case, the synchrotron photons, which are thought to dominate the target radiation[19], are inverse-Compton-scattered to higher energies by the same electron population.

The SSC and synchrotron emission origin scenarios[2] place distinctly different demands on the source acceleration efficiency. Whereas an SSC origin requires electrons with only multi-gigaelectronvolt energies, a synchrotron origin requires an extreme accelerator potentially accelerating beyond petaelectronvolt energies[20] (see Methods). Furthermore, for GRBs to operate as 10^{20} eV cosmic-ray sources, they must achieve extreme acceleration[21]. One key distinguishing characteristic between these two emission origins is that SSC predicts the presence of two bumps in the spectral-energy distribution with their height ratio depending on the energy densities of both the electrons and the magnetic field, whereas a synchrotron model implies only a broad

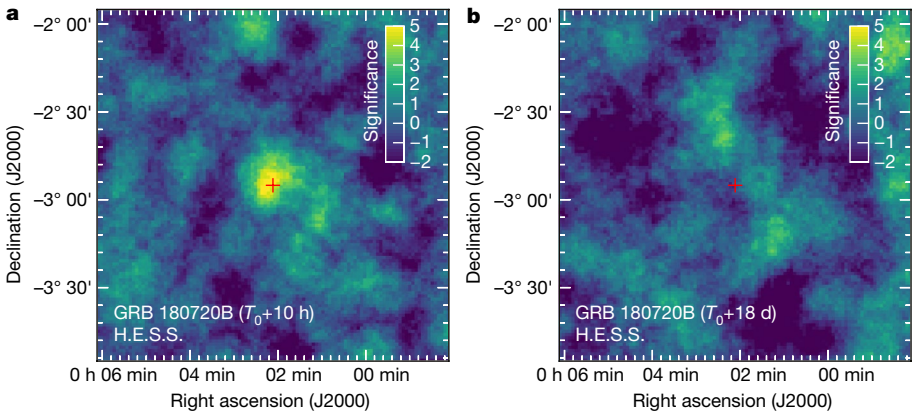


Fig. 2 | Very-high-energy γ-ray image of GRB 180720B. Significance map of GRB 180720B field, as observed by H.E.S.S. **a**, Observation made at $T_0 + 10.1$ h for 2 h. **b**, The same region of the sky, as observed during consecutive nights between $T_0 + 18.4$ d and $T_0 + 24.4$ d. The red cross indicates the position reported by the optical telescope ISON-Castelgrande[12].

single component. A second difference between these processes is the maximum photon energy achievable.

Considering a synchrotron origin of the broadband afterglow energy spectrum, the highest energy for synchrotron emission from electrons in a maximally efficient accelerator is[22] $E_{\rm sync}^{\rm max} = 9\Gamma mc^2/(4\alpha_{\rm F}) \approx 100\Gamma$ MeV (with $\alpha_{\rm F}$ the fine-structure constant and Γ the bulk Lorentz factor of the forward shock). Thus, for electron synchrotron emission to reach energies beyond 100 GeV 10 h after the prompt emission, a late-time Γ in excess of 1,000 appears to be required. Such a scenario is difficult to realize, with robust expectations suggesting a value of $\Gamma \approx 20$ at 10 h (see Methods). Alternatively, circumvention of this synchrotron maximum energy limit is possible for scenarios in which either the coherence length of the magnetic turbulence is very small, or different magnetic-field strengths are present in the acceleration and emission zones, or some non-ideal process is responsible for the particle acceleration (see Methods). Regardless of this challenge, this could explain the similarity in the photon index and level of energy flux of the emission seen both at early times by Fermi-LAT and Swift-XRT and at late times by H.E.S.S. and Swift-XRT (Fig. 1). However, the strong requirements for synchrotron emission to extend up to the VHE regime disfavours such origin and the potential onset of a new SSC component should be considered.

The SSC scenario has the advantage that the emission up to VHE at late times is energetically much more easily achievable[23], leading to the expectation of a new spectral component at VHE. The H.E.S.S. spectral-fit constraints (Extended Data Fig. 1) are consistent with such a possibility within the present uncertainties. Despite this advantage, the potential onset of inverse Compton emission within the Klein–Nishina regime faces challenges (see Methods). Specifically, beyond the γ-ray energy where this sets in, a softer spectral slope and a different brightness evolution of this component[24] are expected. However, interestingly, the presence of synchrotron emission with a hard photon index extending below kiloelectronvolt energies can sufficiently delay the onset of the full Klein–Nishina transition to higher energies (see Methods), beyond that of the VHE detection. The detection of this hard extended synchrotron emission component thus delivers additional supporting evidence for an SSC origin.

This VHE discovery undoubtedly opens a key channel to the understanding of the GRB afterglow phenomena. This measurement proves to be complementary to the VHE-afterglow emission detected in GRB 190829A[25] and the prompt-to-early afterglow emission measured in GRB 190114C by the MAGIC telescopes[26], providing insight into the nature of GRBs and their VHE detectability. We estimate that future instruments, such as the Cherenkov Telescope Array[27], will allow up to three more GRB afterglow detections per year in the VHE domain than previously anticipated (see Methods), considerably improving our understanding of GRBs over a diverse range of timescales.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1743-9.

1. Mészáros, P. Gamma-ray bursts. *Rep. Prog. Phys.* **69**, 2259–2321 (2006).
2. Zhang, B. & Mészáros, P. Gamma-ray bursts: progress, problems & prospects. *Int. J. Mod. Phys. A* **19**, 2385–2472 (2004).
3. Ackermann, M. et al. Fermi-LAT observations of the gamma-ray burst GRB 130427A. *Science* **343**, 42–47 (2014).
4. Piron, F. Gamma-ray bursts at high and very high energies. *C. R. Phys.* **17**, 617–631 (2016).
5. Roberts, O. J. et al. GCN22981 – GRB 180720B: Fermi-GBM observation. *GCN Circulars* https://gcn.gsfc.nasa.gov/gcn3/22981.gcn3 (2018).
6. Siegel, M. H. et al. GCN22973 – GRB 180720B: Swift detection of a burst. *GCN Circulars* https://gcn.gsfc.nasa.gov/gcn3/22973.gcn3 (2018).
7. Malesani, D. et al. GCN22996 – VLT/X-shooter redshift. *GCN Circulars* https://gcn.gsfc.nasa.gov/gcn3/22996.gcn3 (2018).
8. Bissaldi, E. et al. GCN22980 – GRB 180720B: Fermi-LAT detection. *GCN Circulars* https://gcn.gsfc.nasa.gov/gcn3/22980.gcn3 (2018).
9. Levan, A. et al. Gamma-ray burst progenitors. *Space Sci. Rev.* **202**, 33–78 (2016).
10. Evans, P. A. et al. Methods and results of an automatic analysis of a complete sample of Swift-XRT observations of GRBs. *Mon. Not. R. Astron. Soc.* **397**, 1177–1201 (2009).
11. Evans, P. A. et al. An online repository of Swift/XRT light curves of γ-ray bursts. *Astron. Astrophys.* **469**, 379–385 (2007).
12. Schmalz, S. et al. GCN23020 – ISON-Castelgrande observation of GRB 180720B. *GCN Circulars* https://gcn.gsfc.nasa.gov/gcn3/23020.gcn3 (2018).
13. Franceschini, A., Rodighiero, G. & Vaccari, M. Extragalactic optical-infrared background radiation, its time evolution and the cosmic photon-photon opacity. *Astron. Astrophys.* **487**, 837–852 (2008).
14. Ajello, M. et al. A decade of gamma-ray bursts observed by Fermi-LAT: the second GRB catalog. *Astrophys. J.* **878**, 52 (2019).
15. Chevalier, R. A. & Li, Z. Y. Wind interaction models for gamma-ray burst afterglows: the case for two types of progenitors. *Astrophys. J.* **536**, 195–212 (2000).
16. Kumar, P. & Barnoil Duran, R. On the generation of high-energy photons detected by the Fermi Satellite from gamma-ray bursts. *Mon. Not. R. Astron. Soc.* **400**, L75–L79 (2009).
17. Sari, R. & Esin, A. A. On the synchrotron self-Compton emission from relativistic shocks and its implications for gamma-ray burst afterglows. *Astrophys. J.* **548**, 787–799 (2001).
18. Zhang, B. & Mészáros, P. High-energy spectral components in gamma-ray burst afterglows. *Astrophys. J.* **559**, 110–122 (2001).
19. Warren, D. C. et al. Nonlinear particle acceleration and thermal particles in GRB afterglows. *Astrophys. J.* **835**, 248 (2017).
20. Guilbert, P. W., Fabian, A. C. & Rees, M. J. Spectral and variability constraints on compact sources. *Mon. Not. R. Astron. Soc.* **205**, 593–603 (1983).
21. Aharonian, F. A., Belyanin, A. A., Derishev, E. V., Kocharovsky, V. V. & Kocharovsky, V. I. V. Constraints on the extremely high-energy cosmic ray accelerators from classical electrodynamics. *Phys. Rev. D* **66**, 023005 (2002).
22. Aharonian, F. A. TeV gamma rays from BL Lac objects due to synchrotron radiation of extremely high energy protons. *New Astron.* **5**, 377–395 (2000).
23. Warren, D. C., Barkov, M. V., Hirotaka, I., Nagataki, S. & Laskar, T. Synchrotron self-absorption in GRB afterglows: the effects of a thermal electron population. *Mon. Not. R. Astron. Soc.* **480**, 4060 (2018).
24. Nakar, E., Ando, S. & Sari, R. Klein–Nishina effects on optically thin synchrotron and synchrotron self-Compton spectrum. *Astrophys. J.* **703**, 675–691 (2009).
25. de Naurois, M. et al. GRB190829A: Detection of VHE gamma-ray emission with HESS. *The Astronomer's Telegram* **13052** (2019).
26. Mirzoyan, R. First time detection of a GRB at sub-TeV energies; MAGIC detects the GRB 190114C. *The Astronomer's Telegram* **12390** (2019).
27. CTA Consortium. *Science with the Cherenkov Telescope Array* (World Scientific Publishing, 2019).

H. Abdalla[1], R. Adam[2], F. Aharonian[3,4,5], F. Ait Benkhali[3], E. O. Angüner[6], M. Arakawa[7], C. Arcaro[1], C. Armand[8], H. Ashkar[9], M. Backes[1,10], V. Barbosa Martins[11], M. Barnard[1], Y. Becherini[12], D. Berge[11], K. Bernlöhr[3], E. Bissaldi[13,14], R. Blackwell[15], M. Böttcher[1], C. Boisson[16], J. Bolmont[17], S. Bonnefoy[11], J. Bregeon[18], M. Breuhaus[3], F. Brun[9], P. Brun[9], M. Bryan[19], M. Büchele[20], T. Bulik[21], T. Bylund[12], M. Capasso[22], S. Caroff[17], A. Carosi[8], S. Casanova[3,23], M. Cerruti[17,24], T. Chand[1], S. Chandra[1], A. Chen[25], S. Colafrancesco[25,26], M. Curyło[21], I. D. Davids[10], C. Deil[3], J. Devin[27], P. deWilt[15], L. Dirson[26], A. Djannati-Ataï[28], A. Dmytriiev[16], A. Donath[3], V. Doroshenko[22], J. Dyks[29], K. Egberts[30], G. Emery[17], J.-P. Ernenwein[6], S. Eschbach[20], K. Feijen[15], S. Fegan[2], A. Fiasson[8], G. Fontaine[2], S. Funk[20], M. Füßling[11], S. Gabici[28], Y. A. Gallant[18], F. Gaté[8], G. Giavitto[11], L. Giunti[28], D. Glawion[31], J. F. Glicenstein[9], D. Gottschall[22], M.-H. Grondin[27], J. Hahn[3], M. Haupt[11], G. Heinzelmann[26], G. Henri[32], G. Hermann[3], J. A. Hinton[3], W. Hofmann[3], C. Hoischen[30], T. L. Holch[33], M. Holler[34], D. Horns[26], D. Huber[34], H. Iwasaki[7], M. Jamrozy[35], D. Jankowsky[20], F. Jankowsky[31], A. Jardin-Blicq[3], I. Jung-Richardt[20], M. A. Kastendieck[26], K. Katarzyński[36], M. Katsuragawa[37], U. Katz[20], D. Khangulyan[7], B. Khélifi[28], J. King[31], S. Klepser[11], W. Kluźniak[29], Nu. Komin[25], K. Kosack[9], D. Kostunin[11], M. Kreter[1], G. Lamanna[8], A. Lemière[28], M. Lemoine-Goumard[27], J.-P. Lenain[17], E. Leser[11,30], C. Levy[17], T. Lohse[33], I. Lypova[11], J. Mackey[4], J. Majumdar[11], D. Malyshev[22], V. Marandon[3], A. Marcowith[18], A. Mares[27], C. Mariaud[2], G. Martí-Devesa[34], R. Marx[3], G. Maurin[8], P. J. Meintjes[38], A. M. W. Mitchell[3,39], R. Moderski[29], M. Mohamed[31], L. Mohrmann[20],

C. Moore[40], E. Moulin[9], J. Muller[2], T. Murach[11], S. Nakashima[41], M. de Naurois[2], H. Ndiyavala[1], F. Niederwanger[34], J. Niemiec[23], L. Oakes[33], P. O'Brien[40], H. Odaka[42], S. Ohm[11], E. de Ona Wilhelmi[11], M. Ostrowski[35], I. Oya[11], M. Panter[3], R. D. Parsons[3], C. Perennes[17], P.-O. Petrucci[32], B. Peyaud[9], Q. Piel[8], S. Pita[28], V. Poireau[8], A. Priyana Noel[35], D. A. Prokhorov[25], H. Prokoph[11], G. Pühlhofer[22], M. Punch[12,28], A. Quirrenbach[31], S. Raab[20], R. Rauth[34], A. Reimer[34], O. Reimer[34], Q. Remy[18], M. Renaud[18], F. Rieger[3], L. Rinchiuso[9], C. Romoli[3], G. Rowell[15], B. Rudak[29], E. Ruiz-Velasco[3]*, V. Sahakian[43], S. Sailer[3], S. Saito[7], D. A. Sanchez[8], A. Santangelo[22], M. Sasaki[20], R. Schlickeiser[44], F. Schüssler[9], A. Schulz[11], H. M. Schutte[1], U. Schwanke[33], S. Schwemmer[31], M. Seglar-Arroyo[9], M. Senniappan[12], A. S. Seyffert[1], N. Shafi[25], K. Shiningayamwe[10], R. Simoni[19], A. Sinha[28], H. Sol[16], A. Specovius[20], M. Spir-Jacob[28], Ł. Stawarz[35], R. Steenkamp[10], C. Stegmann[11,30], C. Steppa[30], T. Takahashi[37], T. Tavernier[9], A. M. Taylor[11], R. Terrier[28], D. Tiziani[20], M. Tluczykont[26], C. Trichard[2], M. Tsirou[18], N. Tsuji[7], R. Tuffs[3], Y. Uchiyama[7], D. J. van der Walt[1], C. van Eldik[20], C. van Rensburg[1], B. van Soelen[38], G. Vasileiadis[18], J. Veh[20], C. Venter[1], P. Vincent[17], J. Vink[19], H. J. Völk[3], T. Vuillaume[8], Z. Wadiasingh[1], S. J. Wagner[31], R. White[3], A. Wierzcholska[23,31], R. Yang[3], H. Yoneda[37], M. Zacharias[1], R. Zanin[3], A. A. Zdziarski[29], A. Zech[16], A. Ziegler[20], J. Zorn[3], N. Żywucka[1], F. de Palma[45], M. Axelsson[46,47] & O. J. Roberts[48]

[1]Centre for Space Research, North-West University, Potchefstroom, South Africa. [2]Laboratoire Leprince-Ringuet, École Polytechnique, UMR 7638, CNRS/IN2P3, Institut Polytechnique de Paris, Paris, France. [3]Max-Planck-Institut für Kernphysik, Heidelberg, Germany. [4]Dublin Institute for Advanced Studies, Dublin, Ireland. [5]High Energy Astrophysics Laboratory, RAU, Yerevan, Armenia. [6]Aix Marseille Université, CNRS/IN2P3, CPPM, Marseilles, France. [7]Department of Physics, Rikkyo University, Tokyo, Japan. [8]Laboratoire d'Annecy de Physique des Particules, Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, LAPP, Annecy, France. [9]IRFU, CEA, Université Paris-Saclay, Gif-sur-Yvette, France. [10]Department of Physics, University of Namibia, Windhoek, Namibia. [11]DESY, Zeuthen, Germany. [12]Department of Physics and Electrical Engineering, Linnaeus University, Växjö, Sweden. [13]Dipartimento Interateneo di Fisica, Politecnico di Bari, Bari, Italy. [14]Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Bari, Italy. [15]School of Physical Sciences, University of Adelaide, Adelaide, South Australia, Australia. [16]LUTH, Observatoire de Paris, PSL Research University, CNRS, Université Paris Diderot, Paris, France. [17]Laboratoire de Physique Nucléaire et de Hautes Energies, LPNHE, Sorbonne Université, Université Paris Diderot, Sorbonne Paris Cité, CNRS/IN2P3, Paris, France. [18]Laboratoire Univers et Particules de Montpellier, Université Montpellier, CNRS/IN2P3, CC 72, Montpellier, France. [19]GRAPPA, Anton Pannekoek Institute for Astronomy, University of Amsterdam, Amsterdam, The Netherlands. [20]Erlangen Centre for Astroparticle Physics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany. [21]Astronomical Observatory, The University of Warsaw, Warsaw, Poland. [22]Institut für Astronomie und Astrophysik, Universität Tübingen, Tübingen, Germany. [23]Instytut Fizyki Jadrowej PAN, Krakow, Poland. [24]Institut de Ciències del Cosmos (ICC UB), Universitat de Barcelona (IEEC-UB), Barcelona, Spain. [25]School of Physics, University of the Witwatersrand, Johannesburg, South Africa. [26]Institut für Experimentalphysik, Universität Hamburg, Hamburg, Germany. [27]Centre d'Études Nucléaires de Bordeaux Gradignan, Université Bordeaux, CNRS/IN2P3, Gradignan, France. [28]APC, AstroParticule et Cosmologie, Université Paris Diderot, CNRS/IN2P3, CEA/Irfu, Observatoire de Paris, Sorbonne Paris Cité, Paris, France. [29]Nicolaus Copernicus Astronomical Center, Polish Academy of Sciences, Warsaw, Poland. [30]Institut für Physik und Astronomie, Universität Potsdam, Potsdam, Germany. [31]Landessternwarte, Universität Heidelberg, Heidelberg, Germany. [32]Université Grenoble Alpes, CNRS, IPAG, Grenoble, France. [33]Institut für Physik, Humboldt-Universität zu Berlin, Berlin, Germany. [34]Institut für Astro- und Teilchenphysik, Leopold-Franzens-Universität Innsbruck, Innsbruck, Austria. [35]Obserwatorium Astronomiczne, Uniwersytet Jagielloński, Krakow, Poland. [36]Centre for Astronomy, Faculty of Physics, Astronomy and Informatics, Nicolaus Copernicus University, Torun, Poland. [37]Kavli Institute for the Physics and Mathematics of the Universe (WPI), The University of Tokyo Institutes for Advanced Study (UTIAS), The University of Tokyo, Kashiwa, Japan. [38]Department of Physics, University of the Free State, Bloemfontein, South Africa. [39]Physik Institut, Universität Zürich, Zurich, Switzerland. [40]Department of Physics and Astronomy, The University of Leicester, Leicester, UK. [41]RIKEN, Wako, Japan. [42]Department of Physics, The University of Tokyo, Tokyo, Japan. [43]Yerevan Physics Institute, Yerevan, Armenia. [44]Institut für Theoretische Physik, Lehrstuhl IV: Weltraum und Astrophysik, Ruhr-Universität Bochum, Bochum, Germany. [45]Istituto Nazionale di Fisica Nucleare – Sezione di Torino, Turin, Italy. [46]Oskar Klein Centre, Department of Physics, Stockholm University, Stockholm, Sweden. [47]Department of Physics, KTH Royal Institute of Technology, Stockholm, Sweden. [48]Science and Technology Institute, Universities Space Research Association, Huntsville, AL, USA.
*e-mail: contact.hess@hess-experiment.eu

Methods

H.E.S.S. and the GRB follow-up programme

The observations presented in this paper were performed using the H.E.S.S. array of imaging atmospheric Cherenkov telescopes, which is situated at an altitude of 1,800 m in the Khomas highlands of Namibia. H.E.S.S. is sensitive to γ-rays in the energy range from tens of gigaelectronvolts to tens of teraelectronvolts. It consists of five Cherenkov telescopes: four with mirror areas of 108 m^2 placed in a square configuration with a side length of 120 m (CT1–CT4) and a single telescope at the centre (CT5) with a mirror area of 614 m^2. Thanks to its low energy threshold and fast slewing (200° min^{-1})[28], CT5 is well suited for the observation of soft-spectrum transient sources.

H.E.S.S. maintains an active transient-source observation programme, of which GRBs are an important component. To ensure a fast reaction to GRB alerts, H.E.S.S. is connected to the γ-ray coordinates network (GCN)[29], which rapidly distributes alerts and observational information from space and ground-based facilities. The target-of-opportunity observation system in H.E.S.S. performs the selection, filtering and processing of these alerts on the basis of source observability and significance, aiming to trigger on bright, precisely located, nearby bursts. Alerts are followed up in two different observation modes. Observations are triggered in the prompt mode when the GRB position is observable from the H.E.S.S. site at the time that the alert is received. In this case, the observation schedule is interrupted and the array is automatically re-pointed to the GRB location. On the other hand, afterglow observations take place for GRBs that become observable only at a later time; such observations are scheduled manually and are triggered by a burst advocate. This was the case for GRB 180720B, which was observed from T_0 + 10.1 h, when the burst position rose above 45° in elevation (below this elevation GRBs are typically not observed owing to the rapid increase in the energy threshold of the H.E.S.S. telescopes). Re-observations were carried out at T_0 + 18 d, after the end of an intervening moonlight period.

H.E.S.S. data analysis

To reach the lowest possible energies in the analysis presented here, we use only data from the single large telescope (CT5). However, this energy threshold reduction comes at the cost of some angular resolution and sensitivity loss[30]. We present here two hours of observations taken in wobble mode[31], with the pointing direction of the telescope at an offset of 0.5° from the position provided by Swift-BAT[6]. This observation was made at a mean zenith angle of 31.5° for a total live time of 1.8 h.

To ensure that a potential GRB signal is not diminished by an excessive number of statistical trials, the data analysis is subjected to a strict unblinding procedure. The first step in this unblinding is an inspection of the low-level data, as some calibration artefacts can directly lead to the creation of spurious sources in the field of view. Checks are made on the fractional event participation of each camera pixel (to ensure that single faulty pixels do not dominate the events), the pixel pedestal values and the distribution of events within the field of view. Once these checks are completed, with no artefacts found, the event properties are reconstructed using the ImPACT[32,33] maximum-likelihood-based fitting technique. Background cosmic-ray events are rejected using a neural-network-based scheme[34]. The residual background contamination level of the source region (ON and OFF events) and the ratio of the on-source time to the off-source time (α_{Exp}) are then estimated using the ring method for the production of maps and the reflected-region method when performing the spectral extraction[31]. Full analysis and checks are performed using an additional independent calibration and data analysis chain[35], serving as a cross-check of all the results.

The source significance is computed using a maximum-likelihood ratio test based on the number of events coming from the source (ON) and the background (OFF) for a given ratio of on-source to off-source time (α; ref. [36]). For the ring method, the number of ON and OFF events is 544 and 4,740, respectively, and $\alpha_{\mathrm{Exp}} = 0.09$, resulting in a significance of detection of 5.3σ. Similarly the reflected-region method measures 544 ON and 3,998 OFF events and $\alpha_{\mathrm{Exp}} = 0.11$ at a significance of 4.6σ, which is verified by the cross-check analysis, which provided 651 ON and 5,200 OFF events and $\alpha_{\mathrm{Exp}} = 0.10$ with a significance of 4.5σ.

The source morphology is fitted with a two-dimensional likelihood procedure by assuming point-like and Gaussian source models convolved with the expected energy-dependent point spread function (obtained from simulations) and the measured source spectrum. Both source models are proved to be compatible with the morphology of the discovered source, with no statistically significant preference for source extension shown.

Spectral analysis is performed using the forward-folding method[37], which corrects for the limited energy resolution of the single-telescope event reconstruction. The measured source spectrum is obtained by fitting a simple power-law model of the form $F_{\mathrm{obs}}(E) = F_{0,\mathrm{obs}}(E/E_{0,\mathrm{obs}})^{-\gamma_{\mathrm{obs}}}$, where $F_{0,\mathrm{obs}}$ is the flux normalization, γ_{obs} is the photon index and $E_{0,\mathrm{obs}}$ is the reference energy. However, owing to the absorption of the most energetic photons by the extragalactic background light (EBL), the apparent photon index of this source will be somewhat steeper than the intrinsic photon index. The intrinsic spectrum $F_{\mathrm{int}}(E)$ is therefore obtained by fitting the measured spectrum with an attenuated power-law model, $F_{\mathrm{obs}}(E) = F_{\mathrm{int}}(E) \times \mathrm{e}^{-\tau(E,z)} = F_{0,\mathrm{int}}(E/E_{0,\mathrm{int}})^{-\gamma_{\mathrm{int}}} \times \mathrm{e}^{-\tau(E,z)}$, where the last term in the equation corresponds to the EBL absorption coeficient predicted[13] for a redshift of 0.653. The best-fit spectra, together with the spectral points, are shown in Extended Data Fig. 1, and the spectral parameters are summarized in Extended Data Table 1.

Systematic uncertainties in the fitted spectra are determined by accounting for a 15% uncertainty in the reconstructed energy due to possible variations in the measured Cherenkov light yield[38]. The measured energy is systematically shifted by ±15% and the whole spectral-fitting procedure is redone. In addition, short dips in the trigger rate (at the level of 30%) were identified in the data, which can probably be attributed to the presence of high-altitude clouds. To assess the effect of these, the time windows containing such trigger rate features (21.7 min total) were removed from the data and the standard analysis described above was performed on the reduced dataset. From this, we conclude an additional systematic underestimation of 32% and 4.8% in the measured normalization and photon index, respectively. These two sources of systematic uncertainty are considered to be independent and are therefore added in quadrature for the estimation of the total systematic uncertainty.

The intrinsic spectrum was obtained with a chosen EBL model[13]. To determine how this choice influences the results presented in this work, the data were re-analysed using three additional EBL models[39–41], each one employing a different approach to predict the overall EBL level[39]. The absorption coefficient for a redshift of 0.653 within the energy range of the detected emission does not present sizeable deviations between the models considered (Extended Data Fig. 2). When employing these EBL models for the spectral fit, a change of up to 55.3% and 27% was found in the reported normalization and index, respectively. The statistical uncertainty on the fitted spectra remains the biggest source of uncertainty in the results.

Trial correction

Since 2012, H.E.S.S. has performed five additional follow-up observations of well localized GRBs (Swift and Fermi-LAT alerts) using only CT5 (similar to the observations presented here). The significance distribution of this sample (excluding GRB 180720B) is consistent with pure statistical fluctuations. Therefore, the post-trial significance for GRB 180720B can be assessed by accounting for these previously observed GRBs. This results in a post-trial significance of 4.3σ (reflected-region method) and 5.0σ (ring method). As the analysis of GRB 180720B was

performed once under the aforementioned unblinding procedure, no additional trials have been added to the results presented here.

Background systematic effects

Systematic effects on the sky map background (Fig. 2) were determined by measuring the significance distribution when excluding the source region. Although a normal distribution was expected, a width of 1.09 was measured in this significance distribution, therefore adding a slight shift to the reported significance of the ring method (used in the production of sky maps). The corrected significance when accounting for such effects is 4.9σ (4.7σ post-trial). Nonetheless, this measured distribution depends strongly on the parameters of the ring method and should be subject to statistical uncertainties.

Fermi data analysis

The Fermi-GBM data for GRBs are publicly available through the GBM Burst Catalog at HEASARC[42]. For GRB 180720B the available time-tagged events of those detectors having the best viewing angle to the Swift-XRT position—namely, n6, n7, nb and b1—were analysed. Temporally resolved energy-flux data points (Fig. 1) were obtained with the RMfit analysis software[43] by combining time-tagged event data from all four detectors into 256-ms bins in the energy range from 8 keV to 10 MeV.

The analysis of the Fermi-LAT data was performed using the 'Pass8'[44] processed events. We used the P8R3_TRANSIENT010E event class, which is suitable for transient-source analysis, and the corresponding instrument response functions[45]. Events were selected from T_0 to $T_0 + 700$ s in the standard GRB analysis energy range of 100 MeV–100 GeV over a region of 10° around the Swift-XRT localization. Event selection, quality cuts and data analysis were performed with the standard FermiTools[46] software. The source detection over the full duration was determined by a likelihood analysis providing a test-statistic value of TS ≈ 600, which corresponds to a significance of $\sigma \approx 25$ ($\sigma \approx \sqrt{\mathrm{TS}}$).

Because the highest-energy photon detected has an energy of 5 GeV (at $T_0 + 142.4$ s), the temporally resolved energy-flux data points (Fig. 1) were computed in the energy range from 100 MeV to 10 GeV. The analysis model included the Galactic interstellar emission model (gll_iem_v06.fits) and the relative isotropic-diffuse-emission templates provided by the Fermi-LAT collaboration[47], and the normalization of the latter was left free to vary. The spectrum for each bin was fitted by a single power-law model $F(E) = F_0 \times (E/E_0)^{-\gamma}$, with the flux normalization F_0 and the photon index γ as free parameters. As no emission with energies >10 GeV was detected, no additional term was required to account for EBL absorption[13] in the spectra. The temporal decay α_{LAT} was fitted by a power-law model using a least-squares technique applied from $T_0 + 55$ s to $T_0 + 700$ s in order to ensure no contamination of the prompt emission observed by Fermi-GBM and Swift-BAT, obtaining a reduced χ^2 of $\chi^2 = 0.63$ (14 degrees of freedom).

Optical data

The optical data shown in Fig. 1 were compiled from the GCN circulars of observations performed in the r-band by the following instruments: Kanata[48], MITSuME[49], TSHAO[50], MASTER-K[51], MASTER-I, ISON-Castelgrande[12], OSN[52], LCO[53] and KAIT[54]. The reported temporal decay index $\alpha_{\mathrm{optical}}$ was measured from $T_0 + 9{,}642$ s to $T_0 + 3.35 \times 10^5$ s by performing a power-law fit with a χ^2 fitting procedure.

Swift data

The Swift data are publicly available through the Swift online repository[55]. The temporally resolved energy-flux data shown in Fig. 1 were obtained using the Burst Analyser tool[10,11]. The data were rebinned to give a signal-to-noise ratio of 7 and systematic uncertainties were included. The temporal decay reported here (α_{XRT}) was obtained from $T_0 + 2{,}200$ s to $T_0 + 3.05 \times 10^5$ s and corresponds to the fourth break in the light curve, as identified from the fitting procedures of the Swift-XRT tools.

Cherenkov Telescope Array detectability prospects

Considering the Cherenkov Telescope Array (CTA) to be an order of magnitude more sensitive than the H.E.S.S. array implies that it will have the ability to detect energy fluxes ~10 times fainter than that of GRB 180720B at VHE. If the VHE flux equals that detected by Swift-XRT, as suggested by our measurements (Fig. 1), we estimate the occurrence of three GRBs per year above this flux, which will therefore be detectable by CTA (Extended Data Fig. 3). This number could be increased for follow-up observations at earlier times. By assuming a temporal decay value of $\alpha = 1.2$ ($F(t) \propto t^{-\alpha}$) for all the GRB afterglows detected by Swift-XRT[56], an extrapolation of the 11-h energy flux to that expected at 5 h provides a detectability prospect of ~10 GRBs per year at such follow-up delay times. It should be noted, however, that the presence of VHE emission could also be dependent on the GRB environment[57], and this influence was not considered in this estimation.

Bulk Lorentz factor

The bulk Lorentz factor depends on two factors: the released energy and the density of the circumburst medium[58], $\Gamma = \sqrt{E^{\mathrm{iso}}/(Mc^2)}$, where E^{iso} is the equivalent isotropic energy and M is the total mass swept up by the shock. The latter depends on the nature of the circumburst environment: $M = (4\pi/3)R^3 n m_{\mathrm{p}}$ for a homogeneous medium (here n is the medium number density, R is the shock radius and m_{p} is the proton mass) and $M = \dot{M}_* R/\upsilon_*$ for a shock propagating in a constant-velocity wind (here $\dot{M}_*$ and υ_* are the wind mass-loss rate and velocity, respectively). The shock radius depends on the detection time as $R \approx A_{\mathrm{R}} \Gamma^2 tc/(1+z)$, where $A_{\mathrm{R}} = 8$ for a homogeneous medium and $A_{\mathrm{R}} = 4$ for wind environments (c, speed of light in vacuum). Thus, for GRB 180720B ($t = 10$ h, $z = 0.653$, and $E^{\mathrm{iso}} \approx 10^{54}$ erg) one obtains $\Gamma = 15 n_0^{-1/8}$ (here $n_0 = n/(1\,\mathrm{cm}^{-3})$) for a homogeneous medium or $\Gamma = 20 \dot{M}_{*,-5}^{-1/4} \upsilon_{*,3.3}^{1/4}$ (here $\dot{M}_{*,-5} = \dot{M}_*/(10^{-5} M_\odot\ \mathrm{yr}^{-1})$ and $\upsilon_{*,3.3} = \upsilon_*/(2{,}000\ \mathrm{km\ s}^{-1})$) for a wind environment.

Non-thermal process efficiency

The non-thermal process efficiency, $\kappa = t_{\mathrm{dyn}}/t_{\mathrm{cool}}$, depends on the ratio of the shock dynamic timescale, $t_{\mathrm{dyn}} = R/(c\Gamma)$, to the cooling time, t_{cool}, which depends on the radiation mechanism, the density of the target and the energy of the parent particles. For hadronic processes[59], which include proton–proton (pp) and photon–meson ($p\gamma$) channels, the radiation efficiency is $\kappa_{\mathrm{pp}} \approx 10^{-7}[R/(10^{18}\,\mathrm{cm})] n_0$ and $\kappa_{\mathrm{p\gamma}} \approx 3 \times 10^{-4} (\Gamma/20)^2 [R/(10^{18}\,\mathrm{cm})](\kappa/10^{-2})[E_{\mathrm{t}}/(1\,\mathrm{keV})]^{-1} n_0$ (here κ is the total radiative efficiency and E_{t} is the peak frequency of the soft-emission component). These low efficiencies favour the electromagnetic processes[60]. The efficiency of the synchrotron channel for the emission detected in the VHE band, $E \approx 100$ GeV, is $\kappa_{\mathrm{sync}} = 5 \times 10^7 (m_{\mathrm{e}}/m)^{5/2} [R/(10^{18}\,\mathrm{cm})](\eta_{\mathrm{B}}/0.1)^{3/4} \eta_{\mathrm{turb}}^{-1/2} n_0^{3/4}$, where m_{e} and m are the masses of the electron and the emitting particle, respectively, η_{B} is a fraction of the internal energy contained in the magnetic field and $\eta_{\mathrm{turb}} = \max(1, R_{\mathrm{Lar}}^{\mathrm{nr}}/\lambda_{\mathrm{cor}})$ defines the shift of the peak energy if a charged particle interacts with a turbulent magnetic field[61] (here $R_{\mathrm{Lar}}^{\mathrm{nr}}$ is the non-relativistic Larmor radius and λ_{cor} is the magnetic-field correlation length). If the inverse Compton scattering proceeds in the Thomson regime, then $\kappa_{\mathrm{IC}} = 3(\Gamma/20)[\kappa/(10^{-2})]\,[R/(10^{18}\,\mathrm{cm})][E_{\mathrm{t}}/(1\,\mathrm{keV})]^{-1/2} n_0$. Efficiencies larger than 1 indicate that particle cooling occurs faster than the source dynamical timescale and is therefore highly efficient.

Synchrotron emission

Synchrotron emission is characterized by the highest radiation efficiency, but this emission component peaks below the limiting energy of $E_{\mathrm{sync}}^{\mathrm{max}} = 100 \Gamma (m/m_{\mathrm{e}}) \eta_{\mathrm{turb}} (B_{\mathrm{em}}/B_{\mathrm{ac}})(E_{\mathrm{ac}}/B_{\mathrm{ac}})$MeV. Here B_{em} and B_{ac} are the magnetic-field strengths at the emitter and accelerator sites, respectively. The accelerating electric field, E_{ac}, is smaller than the magnetic field, $E_{\mathrm{ac}} < B_{\mathrm{ac}}$, if the particle acceleration proceeds in ideal magnetohydrodynamic flows[21]. Thus, the production of VHE γ-rays via electron synchrotron emission requires a large Lorentz factor, $\Gamma > 10^3$, a very-small-scale magnetic turbulence, $\lambda_{\mathrm{cor}} < 10^{-2} R_{\mathrm{Lar}}^{\mathrm{nr}}$, a large change

of the magnetic-field strength, $B_{em} > 10^2 B_{ac}$, particle acceleration to operate in the non-ideal magnetohydrodynamic regime, or a combination of these factors. Proton synchrotron emission alleviates these requirements, but at the expense of a significantly lower radiation efficiency. Whereas proton synchrotron emission dominates over other hadronic radiation processes in terms of efficiency[18], its efficiency is still considerably smaller than that of electrons. Thus proton synchrotron emission is expected to give rise only to a subdominant emission component within the VHE band.

Energy of particles emitting in the VHE regime

The energy of particles emitting in the VHE regime depends on the dominant radiation mechanism and the properties of the ejecta. In the case of a synchrotron origin scenario, the particle energy is determined by three important factors: the shock Lorentz factor, the strength of the magnetic field and the turbulence scale. The first factor, $\Gamma \approx 20$, is relatively well defined by the epoch of the H.E.S.S. observation, but the magnetic-field strength and the possibility of small-scale turbulence remain highly uncertain. The internal energy density, $\sim 0.1(\Gamma/20)^2 n_0\,\mathrm{J\,m^{-3}}$, suggests that a Gauss-strength magnetic field is expected for the case of energy equipartition between the magnetic field and particles. We note, however, that substantially smaller plasma magnetization is reported in the literature[62], corresponding to weaker magnetic fields by several orders of magnitude. Assuming that synchrotron emission beyond the 100 MeV energy limit in the co-moving frame can be achieved, the energy of the emitting electrons can be estimated as $E_e \approx 4[E/(100\,\mathrm{keV})]^{1/2}(\Gamma/20)^{-1/2}[B/(0.1\,\mathrm{G})]^{-1/2}\eta_{turb}^{-1/2}$ TeV. The production of 100-GeV γ-rays through a synchrotron scenario therefore requires electrons of ultrahigh-energy, $E_e \approx 4$ PeV, unless a configuration with a very-small-scale turbulence is present. The energy of particles that provide the dominant contribution to the inverse Compton emission depends strongly on the spectrum of the target photons and the bulk Lorentz factor. An electron with energy E_e up-scatters a target photon with energy E_t to an energy of $\min\{E_t[E_e/(m_e c^2)]^2, \Gamma E_e\}$. For target photons detected in the X-ray energy band, $E_t \approx 1$ keV, electrons with energy of $E_e \approx 10$ GeV, which in the laboratory frame have an energy of hundreds of gigaelectectronvolts, can produce γ-rays that are detected in the VHE band.

Target photons

Target photons of very different energies can be up-scattered to γ-rays of the same energy. This can be of particular relevance for VHE γ-rays detected from GRBs, where both the target photons and non-thermal electrons probably have broad energy distributions. Assuming a power-law distribution for the target photon flux, $\mathrm{d}n/\mathrm{d}E_t \propto E_t^{-\gamma}$, and for electrons, $\mathrm{d}n_e/\mathrm{d}E_e \propto E_e^{-\gamma_e}$, one finds that the relative contribution to the γ-ray emission depends on the electron energy as $\propto\{1 - [E/(E_e\Gamma)]\}E_e^{2\gamma-\gamma_e}$. For simplicity, just a single high-energy term in the cross-section was accounted for (resulting in the factor $1 - [E/(E_e\Gamma)]$), which is sufficient for a qualitative study. However, the obtained dependence shows that for a reasonable range of photon and electron indices, $1.5 < \gamma, \gamma_e < 3$, the highest-available-energy electrons may provide an important contribution to the γ-ray energy band by up-scattering photons with energies within the infrared-to-ultraviolet range.

Klein–Nishina cutoff

The Klein–Nishina cutoff is a substantial reduction of the Compton cross-section that occurs when $E_e E_t \gtrsim \Gamma m_e^2 c^4$, where E_e and E_t are the electron and target photon energies in the co-moving frame and the laboratory system, respectively. This results in a softening of the γ-ray spectrum that occurs for $E \gtrsim 50(\Gamma/20)^2[E_t/(1\,\mathrm{keV})]^{-1}$ GeV. Because typically the GRB synchrotron spectral-energy distribution peaks in the kiloelectronvolt band, the inverse Compton component detected at late afterglow phases may be affected by the Klein–Nishina cutoff, resulting in reduced fluxes and steeper spectra. This may appear to contradict the relatively hard intrinsic spectral index of $\gamma_{int} \approx 1.6$ inferred from the H.E.S.S. measurement. There are, however, two effects that can result in spectral hardening at energies around the cutoff: (i) the up-scattering of low-energy infrared-to-ultraviolet photons, which give an intrinsic VHE component with the same slope as that seen in the hard-X-ray band and (ii) the hardness of the electron spectrum at gigaelectronvolt energies, where adiabatic losses probably render the electron spectrum hard. The search for consistency within this framework of the hard VHE spectrum with the SSC scenario, however, requires detailed dedicated simulations, which are beyond the scope of this observational paper.

Data and code availability

The raw H.E.S.S. data and the code used in this study are not public, but belong to the H.E.S.S. collaboration. All derived higher-level data that are shown in the plots will be made available on the H.E.S.S. collaboration's website upon publication of this study. Data and analysis code from the Fermi-GBM and LAT instruments are publicly available. Links to the data and software are provided in the Methods section. This work also made use of data supplied by the UK Swift Science Data Centre at the University of Leicester (http://www.swift.ac.uk/archive/).

28. Hofverberg, P. et al. Commissioning and initial performance of the H.E.S.S. II drive system. In *Proc. of the 33rd International Cosmic Ray Conference (ICRC 2013)*, 3092 (Curran Associates, 2013).
29. Bathelmy, S. *GCN: The gamma-ray burst coordinates network* http:/gcn.gsfc.nasa.gov (2019).
30. Holler, M. et al. Observations of the Crab Nebula with H.E.S.S. Phase II. *PoS Proc. Sci. ICRC2015*, **847** (2016).
31. Berge, D., Funk, S. & Hinton, J. Background modelling in very-high-energy gamma-ray astronomy. *Astron. Astrophys*. **466**, 1219–1229 (2007).
32. Parsons, R. D. & Hinton, J. A. A Monte Carlo template based analysis for air-Cherenkov arrays. *Astropart. Phys*. **56**, 26–34 (2014).
33. Parsons, R. D., Murach, T. & Gajdus, M. HESS II data analysis with ImPACT. *PoS Proc. Sci. ICRC2015*, **826** (2015).
34. Murach, T., Gajdus, M. & Parsons, R. D. A neural network-based monoscopic reconstruction algorithm for H.E.S.S. II. *PoS Proc. Sci. ICRC2015*, **1022** (2015).
35. de Naurois, M. & Rolland, L. A high performance likelihood reconstruction of γ-rays for imaging atmospheric Cherenkov telescopes. *Astropart. Phys*. **32**, 231–252 (2009).
36. Li, T. P. & Ma, Y. Q. Analysis methods for results in gamma-ray astronomy. *Astrophys. J*. **272**, 317–324 (1983).
37. Piron, F. et al. Temporal and spectral gamma-ray properties of Mkn 421 above 250 GeV from CAT observations between 1996 and 2000. *Astron. Astrophys*. **374**, 895–906 (2001).
38. Abdalla, H. et al. Gamma-ray blazar spectra with H.E.S.S. II mono analysis: the case of PKS 2155-304 and PG 1553+113. *Astron. Astrophys*. **600**, A89 (2017).
39. Domínguez, A. et al. Extragalactic background light inferred from AEGIS galaxy-SED-type fractions. *Mon. Not. R. Astron. Soc*. **410**, 2556–2578 (2011).
40. Finke, J. D., Razzaque, S. & Dermer, C. D. Modeling the extragalactic background light from stars and dust. *Astrophys. J*. **712**, 238–249 (2010).
41. Gilmore, R. C. et al. Semi-analytic modelling of the extragalactic background light and consequences for extragalactic gamma-ray spectra. *Mon. Not. R. Astron. Soc*. **422**, 3189–3207 (2012).
42. *FERMIGBRST – Fermi GBM Burst Catalog* https://heasarc.gsfc.nasa.gov/W3Browse/fermi/fermigbrst.html (2019).
43. *GBM Software Tools* https://fermi.gsfc.nasa.gov/ssc/data/analysis/rmfit/ (2019).
44. Atwood, W. et al. Pass 8: toward the full realization of the Fermi-LAT scientific potential. In *2012 Fermi Symposium proceedings, eConf C121028* (2013); preprint at https://arxiv.org/abs/1303.3514.
45. *Fermi LAT Performance* http://www.slac.stanford.edu/exp/glast/groups/canda/lat_Performance.htm (2019).
46. *Fermitools-conda-recipe* https://github.com/fermi-lat/Fermitools-conda/ (2019).
47. Acero, F. et al. Development of the model of galactic interstellar emission for standard point-source analysis of Fermi Large Area Telescope data. *Astrophys. J. Suppl. Ser*. **223**, 26 (2016).
48. Sasada, M. et al. GCN22977 – Kanata observation. *GCN Circulars* https://gcn.gsfc.nasa.gov/gcn3/22977.gcn3 (2018).
49. Itoh, R. et al. GCN22983 – MITSuME Akeno observation. *GCN Circulars* https://gcn.gsfc.nasa.gov/gcn3/22983.gcn3 (2018).
50. Reva, I. et al. GCN22979 – TSHAO observation. *GCN Circulars* https://gcn.gsfc.nasa.gov/gcn3/22979.gcn3 (2018).
51. Lipunov, V. et al. GCN23023 – MASTER observation. *GCN Circulars* https://gcn.gsfc.nasa.gov/gcn3/23023.gcn3 (2018).
52. Kann, D. et al. GCN22985 – OSN observation. *GCN Circulars* https://gcn.gsfc.nasa.gov/gcn3/22985.gcn3 (2018).
53. Martone, R. et al. GCN22976 – LCO Haleaka observation of GRB 180720B. *GCN Circulars* https://gcn.gsfc.nasa.gov/gcn3/22976.gcn3 (2018).
54. Zheng, W. et al. GCN23033 – KAIT observation of GRB 180720B. *GCN Circulars* https://gcn.gsfc.nasa.gov/gcn3/23033.gcn3 (2018).

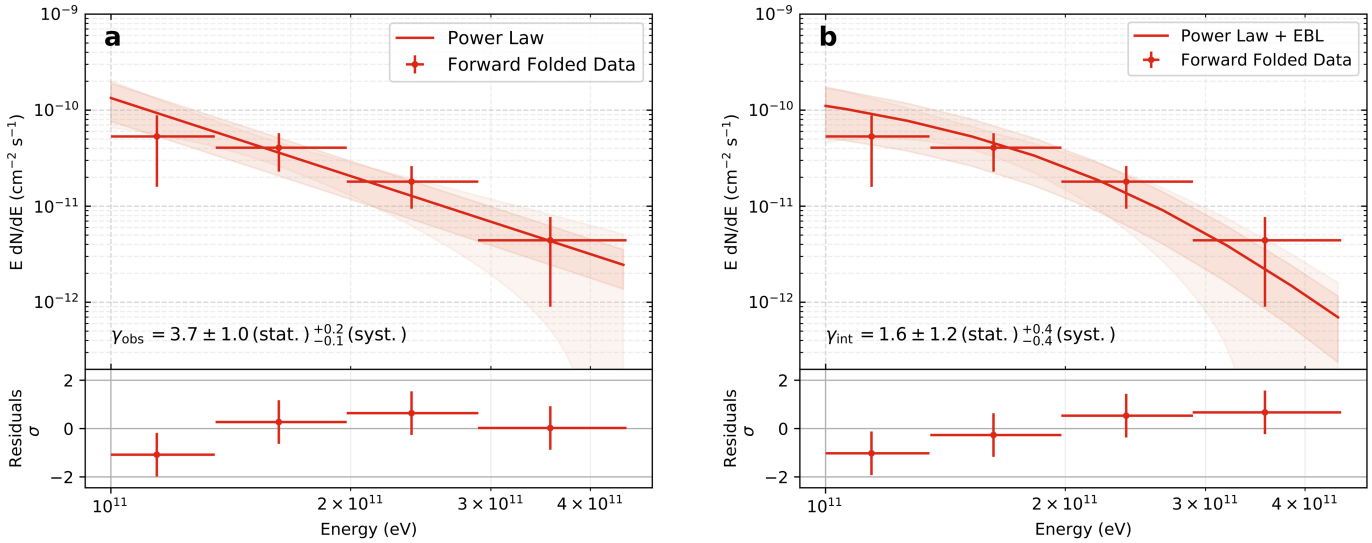
55. The Swift Burst Analyser – GRB *180720B* https://www.swift.ac.uk/burst_analyser/00848890/ (2018).
56. Nousek, J. A. et al. Evidence for a canonical GRB afterglow light curve in the Swift/XRT data. *Astrophys. J.* **642**, 389–400 (2006).
57. Vurm, I. & Beloborodov, A. M. On the prospects of gamma-ray burst detection in the TeV band. *Astrophys. J.* **846**, 152 (2017).
58. Blandford, R. D. & McKee, C. F. Fluid dynamics of relativistic blast waves. *Phys. Fluids* **19**, 1130–1138 (1976).
59. Aharonian, F. A. *Very High Energy Cosmic Gamma Radiation: A Crucial Window on the Extreme Universe* (World Scientific Publishing, 2004).
60. Blumenthal, G. R. & Gould, R. J. Bremsstrahlung, synchrotron radiation, and Compton scattering of high-energy electrons traversing dilute gases. *Rev. Mod. Phys.* **42**, 237–270 (1970).
61. Kelner, S. R., Aharonian, F. A. & Khangulyan, D. On the jitter radiation. *Astrophys. J.* **774**, 61 (2013).
62. Santana, R., Barniol, D. & Kumar, P. Magnetic fields in relativistic collisionless shocks. *Astrophys. J.* **785**, 29 (2014).

Acknowledgements We acknowledge the support of the Namibian authorities and of the University of Namibia in facilitating the construction and operation of H.E.S.S. We also acknowledge support by the German Ministry for Education and Research (BMBF); the Max Planck Society; the German Research Foundation (DFG); the Helmholtz Association; the Alexander von Humboldt Foundation; the French Ministry of Higher Education, Research and Innovation; the Centre national de la recherche scientifique (CNRS/IN2P3 and CNRS/INSU); the Commissariat a l'energie atomique et aux energies alternatives (CEA); the UK Science and Technology Facilities Council (STFC); the Knut and Alice Wallenberg Foundation; the National Science Centre, Poland, through grant number 2016/22/M/ST9/00382; the South African Department of Science and Technology and National Research Foundation; the University of Namibia; the National Commission on Research, Science & Technology of Namibia (NCRST); the Austrian Federal Ministry of Education, Science and Research; the Austrian Science Fund (FWF); the Australian Research Council (ARC); the Japan Society for the Promotion of Science; and the University of Amsterdam. We appreciate the work of the technical support staff in Berlin, Zeuthen, Heidelberg, Palaiseau, Paris, Saclay, Tübingen and in Namibia for the construction and operation of the equipment. This work benefited from services provided by the H.E.S.S. Virtual Organisation, supported by the national resource providers of the EGI Federation. The Fermi-LAT Collaboration acknowledges support for LAT development, operation and data analysis from NASA, the US Department of Energy (DOE), CEA/Irfu and IN2P3/CNRS (France), ASI and INFN (Italy), MEXT, KEK, JAXA (Japan), the K. A. Wallenberg Foundation, the Swedish Research Council and the National Space Board (Sweden). Science analysis support in the operations phase from INAF (Italy) and CNES (France) is also acknowledged. This work was performed in part under DOE contract DE-AC02-76SF00515.

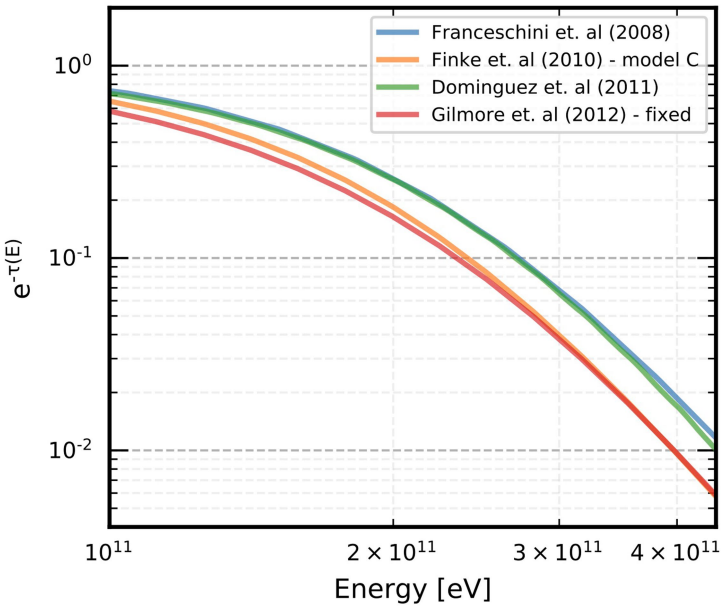
Author contributions R.D.P., Q.P. and E.R.-V. analysed and interpreted the H.E.S.S. data and prepared the manuscript. E.B. analysed and interpreted the Fermi data and prepared the manuscript. C.H. implemented the response system for the GRBs follow-up of H.E.S.S. A.M.T., F.A. and D. Khangulyan helped to interpret the results and prepare the manuscript. The entire H.E.S.S. collaboration contributed to the publication with involvement at various stages, from the design, construction and operation of the instrument to the development and maintenance of all software for data handling, data reduction and data analysis. All authors reviewed, discussed and commented on the present results and the manuscript.

Competing interests The authors declare no competing interests.

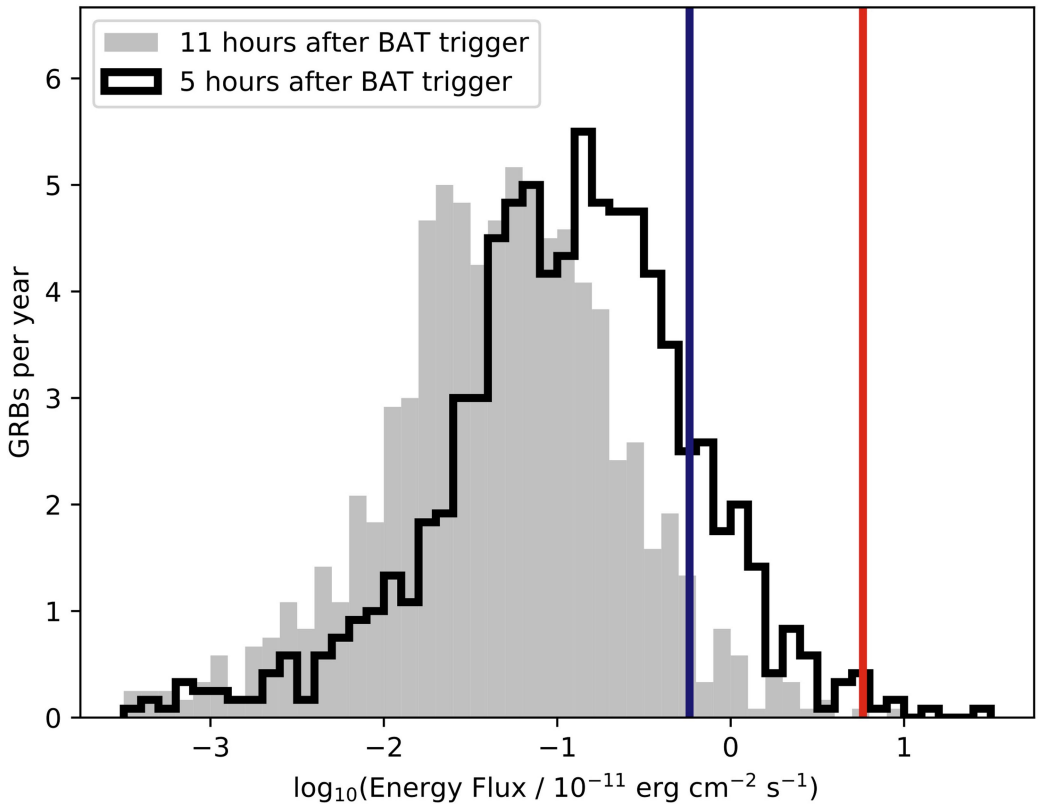
Additional information
Correspondence and requests for materials should be addressed to E.R.-V.
Peer review information *Nature* thanks Gus Sinnis and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | VHE spectral fit of GRB 180720B. H.E.S.S. spectral fit to the measured emission in the energy range 100–440 GeV. **a**, Fit using a simple power-law model (with photon index γ_{obs}). **b**, Fit with a power-law model (with photon index γ_{int}) with EBL attenuation for a source at $z = 0.653$ (ref. [13]). In both cases the residual data points with 1σ uncertainties are obtained from the forward-folded method. The shaded areas show the statistical and systematic uncertainties in each fit (1σ confidence level). The bottom panels show the significance of the residuals between the fitted model and the data points.



Extended Data Fig. 2 | EBL absorption coefficient. Absorption coefficient $e^{-\tau(E)}$ for a source emitting at a redshift of 0.653. The values are shown in the energy range of the detected emission of GRB 180720B (100–440 GeV) for the four EBL models considered[13,39–41].



Extended Data Fig. 3 | CTA detectability prospects. Energy-flux distribution at 11 h and 5 h after the Swift-BAT trigger for all the GRBs observed by Swift-XRT per year. The blue vertical line shows the expected sensitivity of CTA, assuming the detection of fluxes 10 times fainter than that of GRB 180720B. The energy flux of GRB 180720B is indicated by the red vertical line.

Extended Data Table 1 | VHE spectral information from GRB 180720B

Spectral model	F_0 [$\times 10^{-10}$ cm^{-2} s^{-1} TeV^{-1}]	γ	E_0 [TeV]
$F_0 \times \left(\frac{E}{E_0}\right)^{-\gamma}$	$2.71 \pm 0.74^{+1.43}_{-1.16}$	$3.7 \pm 1.0^{+0.2}_{-0.1}$	0.154
$F_0 \times \left(\frac{E}{E_0}\right)^{-\gamma} \times e^{-\tau(z,E)}$	$7.52 \pm 2.03^{+4.53}_{-3.84}$	$1.6 \pm 1.2^{+0.4}_{-0.4}$	0.154
$F_0 \times \left(\frac{E}{E_0}\right)^{-2} \times e^{-\tau(z,E)}$	$16.12 \pm 4.37^{+10.59}_{-9.25}$	2.0 [Fixed]	0.105

Spectral parameters of the fits to the H.E.S.S. observed emission in the energy range 100–440 GeV. The intrinsic spectrum with $\gamma = 2.0$ (third row) is provided as a reference to the Fermi-LAT mean photon index detected in several other GRBs at high energies[14]. All reported uncertainties are statistical and systematic, in that order.

Large electrocaloric effects in oxide multilayer capacitors over a wide temperature range

https://doi.org/10.1038/s41586-019-1634-0

Received: 27 September 2018

Accepted: 18 July 2019

Published online: 9 October 2019

B. Nair[1], T. Usui[2], S. Crossley[1], S. Kurdi[1], G. G. Guzmán-Verri[1,3,4], X. Moya[1]*, S. Hirose[2]* & N. D. Mathur[1]*

Heat pumps based on magnetocaloric and electrocaloric working bodies—in which entropic phase transitions are driven by changes of magnetic and electric field, respectively—use displaceable fluids to establish relatively large temperature spans between loads to be cooled and heat sinks[1,2]. However, the performance of prototypes is limited because practical magnetocaloric working bodies driven by permanent magnets[3–5] and electrocaloric working bodies driven by voltage[6–16] display temperature changes of less than 3 kelvin. Here we show that high-quality multilayer capacitors of $PbSc_{0.5}Ta_{0.5}O_3$ display large electrocaloric effects over a wide range of starting temperatures when the first-order ferroelectric phase transition is driven supercritically (as verified by Landau theory) above the Curie temperature of 290 kelvin by electric fields of 29.0 volts per micrometre. Changes of temperature in the large central area of the capacitor peak at 5.5 kelvin near room temperature and exceed 3 kelvin for starting temperatures that span 176 kelvin (complete thermalization would reduce these values from 5.5 to 3.3 kelvin and from 176 to 73 kelvin). If magnetocaloric working bodies were to be replaced with multilayer capacitors of $PbSc_{0.5}Ta_{0.5}O_3$, then the established design principles behind magnetocaloric heat pumps could be repurposed for better performance without bulky and expensive permanent magnets.

The development of electrocaloric (EC) cooling devices in past[6–8] and recent[9–16] times continues to lag the highly developed activity on near-room-temperature magnetocaloric (MC) cooling devices[1,3]. In parallel, the recent development of devices[17] in which uniaxial stress drives elastocaloric materials[2] is complemented by the nascent development of devices in which isotropic stress drives barocaloric materials[2]. Thermal changes are particularly large in these two types of mechanocaloric material[2,18], but it is difficult to adapt the highly evolved MC prototypes to use mechanocaloric materials instead.

These MC prototypes[3] typically use permanent magnets to address beds of commercial-grade Gd spheres, whose adiabatic temperature change of $|\Delta T| \approx 2.5$ K (ref. [4]) drives heat exchange with a fluid, permitting heat to be pumped over much larger temperature spans. These temperature spans are established either along the fluid alone (passive regeneration)[1] or along the bed and the fluid together (active regeneration)[3], such that heat is absorbed from the load at the cold end, and dumped to the sink at the hot end. MC effects larger than 2.5 K can be achieved in Gd by increasing the magnetic field and reducing the demagnetizing factor, but these modifications would be challenging to use in practical applications (see Supplementary Note 1). EC effects larger than 2.5 K have not previously been demonstrated in macroscopic bodies (bulk samples and MLCs) near room temperature if one requires unambiguous evidence in the form of directly measured temperature change (Supplementary Note 2).

Bulk EC ceramics such as those used in EC cooling devices[6–8,10] are no thicker than 0.1–0.5 mm, to avoid unduly compromising the breakdown field[2], but the largest applied fields yield at best a directly measured value of $|\Delta T| \approx 2.2$ K near room temperature (in $PbSc_{0.5}Ta_{0.5}O_3$; that is, PST)[6–8,19]. A larger temperature change may be achieved in order-of-magnitude thinner EC films[20], but there is insufficient active material for applications (and therefore an innovative device[15] based on the electrostatic actuation of a flexible polymer bilayer would struggle to continuously cool a macroscopic object). However, an assembly of EC films in the form of a multilayer capacitor (MLC) represents a viable working body that is macroscopic[21]. Although MLCs have now been exploited in several EC cooling devices[9,11–14,16], directly measured temperature jumps $|\Delta T_j|$ have been limited to 2.2 K near room temperature (in MLCs based on a polymer)[9] and 2.7 K near 380 K (in MLCs based on $0.9Pb(Mg_{1/3}Nb_{2/3})O_3–0.1PbTiO_3$)[22]. These highly adiabatic temperature jumps arise within the active area (where interdigitated electrodes overlap) following rapid thermalization between active and inactive layers, and are typically measured at the face centres of MLCs.

Here we describe high-quality MLCs based on the well-known EC material PST[6–8,23,24], which is paraelectric near and above room

[1]Department of Materials Science, University of Cambridge, Cambridge, UK. [2]Murata Manufacturing Co., Ltd., Kyoto, Japan. [3]Centro de Investigación en Ciencia e Ingeniería de Materiales, Universidad de Costa Rica, San José, Costa Rica. [4]Escuela de Física, Universidad de Costa Rica, San José, Costa Rica. *e-mail: xm212@cam.ac.uk; h_sakyo@murata.com; ndm12@cam.ac.uk

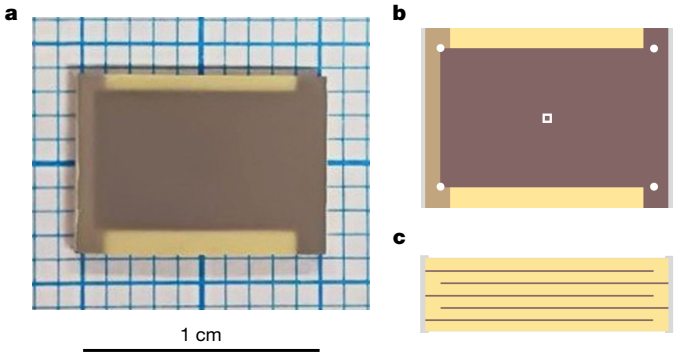


Fig. 1 | MLC structure. a, Optical image, and **b**, schematic illustration, showing the uppermost inner electrode (dark brown), part of the next inner electrode (light brown), the two outer electrodes (grey) and through-thickness PST (yellow). Image contrast was optimized to distinguish each region. The MLC shown here is equivalent to MLC1. In **b**, the white square denotes the face centre (300 μm × 300 μm), and the four white dots define the corners of a rectangle that represents the active area in which the inner electrodes overlap. **c**, Cross-sectional schematic, showing six rather than all 21 layers of PST.

temperature. On cooling, there is a broad transition to a relaxor state (with glassy ferroelectric order) if the B-site cation order is low, and there is a sharp first-order phase transition to a ferroelectric state if the B-site cation order is high. We used highly ordered PST to access large EC effects associated with the latent heat of the first-order transition, which increases with increasing B-site cation order[25,26] as expected for a first-order transition when varying the degree of disorder[27]. By driving this transition supercritically, we accessed about 1.5 times the entropy associated with this latent heat. To do this while avoiding electrical breakdown and leakage-induced Joule heating required fine grains of similar size, a low density of physical and chemical defects, and no discernible impurity phases. Our sintering and annealing conditions were optimized to yield all of these properties, as well as high B-site cation order. We copied the MLC geometry that we previously used for MLCs of $0.9Pb(Mg_{1/3}Nb_{2/3})O_3$–$0.1PbTiO_3$, in which 19 active layers yielded a larger value of $|\Delta T_j|$ than 14 layers[22] or 49 layers[28], and in which the layer thickness of about 40 μm falls in a broad range for which the breakdown field is maximized. However, we increased the active area at the expense of the surrounding inactive area to generate more EC heat and reduce internal thermalization, while retaining sufficient inactive area to suppress cracking by suppressing piezoelectricity. The inactive area also hinders breakdown by suppressing electrical discharge between the inner electrodes.

EC effects were directly measured with an infrared camera whose field of view covered most of the MLC face. The resulting values of $|\Delta T_j|$ reported here describe the face centre and were corroborated by using a thermocouple. As a result of driving the first-order ferroelectric phase transition supercritically without electrical breakdown, we were able to achieve large EC effects over starting temperatures whose wide range would traditionally be associated with smaller EC effects in relaxors[2]. Specifically, our measured changes of temperature peaked at $|\Delta T_j| \approx 5.5$ K near room temperature and exceeded $|\Delta T_j| \approx 3$ K for starting temperatures that spanned 176 K in the range 294–470 K. As we show in this paper, the height and width of this peak compare favourably with respect to the MC effects[4] that can be driven in a macroscopic volume of commercial-grade Gd using bulky and expensive permanent magnets, even in the limit of complete internal MLC thermalization. Our EC capacitors should therefore permit MC prototype

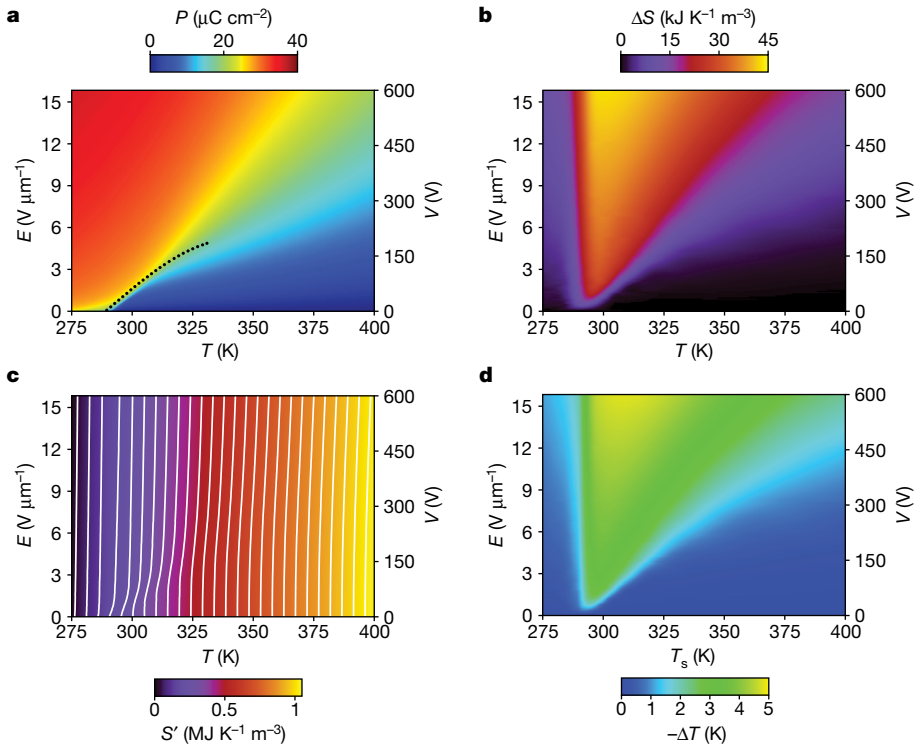


Fig. 2 | Indirect EC measurements. a, Polarization $P(T,E)$ constructed from upper branches of 385 unipolar $P(E)$ plots measured isothermally every 0.35 K on warming (dotted line shows phase boundary with critical endpoint). **b**, The reversible isothermal entropy change $\Delta S(T,E) > 0$ due to the removal of field E, as determined from **a**. **c**, Entropy $S'(T,E)$ obtained by subtracting EC entropy change $\Delta S(T,E) > 0$ from zero-field entropy $S'(T) = S(T) - S(270\,\mathrm{K})$ (Supplementary Fig. 3b), white adiabatic contour separation is about 4.8 K. **d**, Reversible adiabatic temperature change $\Delta T(T_s,E) < 0$ due to the removal of field E at starting temperature T_s, as determined from **c**. Data for MLC1. Selected $P(E)$ plots in Supplementary Note 7. Selected constant-field cross-sections in Supplementary Note 8. Abscissae exclude both extremes of the 270–405 K measurement temperature range, for reliable evaluation of $(\partial P/\partial T)_E$.

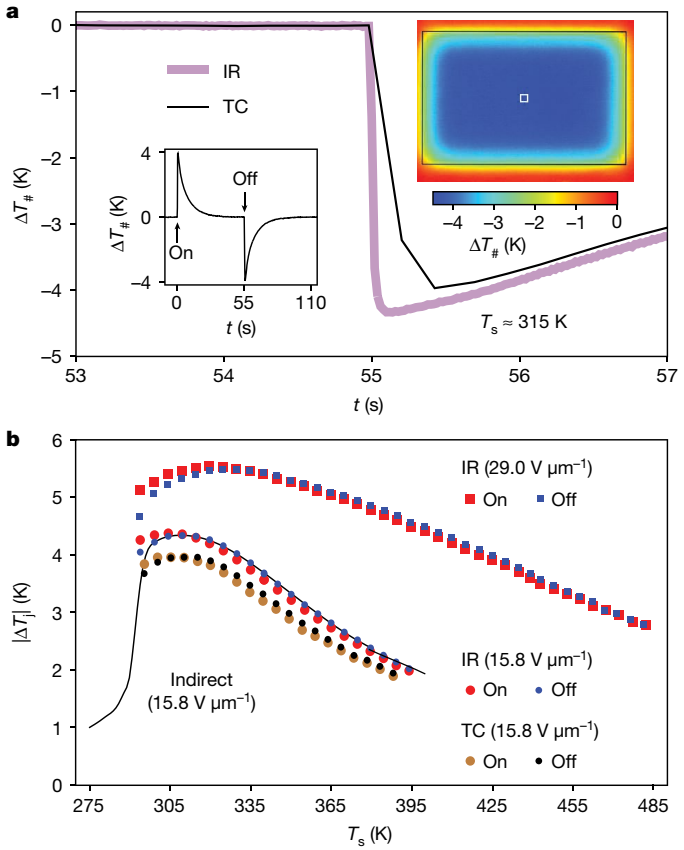


Fig. 3 | Direct EC measurements. a, Left inset shows temperature change $\Delta T_{\#}$ versus time t when a field of 15.8 V μm^{-1} was switched on and off. Data were measured at the MLC face centre with the thermocouple, starting at temperature $T_s \approx 315$ K. The main panel shows a detail of this inset (black data), as well as the corresponding infrared camera measurements for the 300 μm × 300 μm face centre (purple data). Right inset shows an infrared image of the MLC face at $t \approx 55.1$ s, when the face centre (white square overlay) and surrounding area were coldest (dark blue). The black rectangular overlay outlines the active area. IR, infrared; TC, thermocouple. **b**, For different values of T_s, we plot temperature jump magnitude $|\Delta T_j|$ when a field of 15.8 V μm^{-1} or 29.0 V μm^{-1} was switched on and later off. Data were measured at the MLC face centre using the infrared camera or thermocouple. Black line shows $0.90|\Delta T|$ versus T_s for 15.8 V μm^{-1}, identified using $|\Delta T|$ from the indirect method (Fig. 2d) after scaling by 0.90 to achieve a least-squares fit with field-off values of $|\Delta T_j|$ (blue circles). All data for MLC1. Similar data for five similar MLCs appear in Supplementary Fig. 8.

design principles[3] to be repurposed for improved performance, without permanent magnets.

MLC fabrication, characterization and measurement are described in the Methods; the challenges of fabrication are discussed in Supplementary Note 3; and our direct EC measurement set-ups are shown in Supplementary Note 4. A plan-view optical image (Fig. 1a) and schematic (Fig. 1b) accompany a cross-sectional schematic (Fig. 1c) showing six rather than all 21 layers of PST. The 19 active layers, of average thickness 37.9 μm, were electrically addressed by interdigitated inner electrodes of Pt (about 2 μm thick), the active area was approximately 49 mm^2, and the active volume occupied 54% of the total MLC volume (10.45 mm × 7.43 mm × 0.84 mm).

X-ray diffraction revealed a high degree of B-site cation order (order parameter $S_{111} \approx 0.96$; see Supplementary Note 5). This represents a substantial improvement with respect to MLCs of PST ($S_{111} \approx 0.6$–0.7), for which $|\Delta T| \approx 2.4$ K in the active volume[24] was identified via a type of correction that can be prone to overestimation[22]. Zero-field calorimetry revealed a Curie temperature and latent heat ($T_C \approx 290$ K and $|Q_0| \approx 10.2$ MJ m^{-3}, inset of Supplementary Fig. 3a) that differ only slightly from the corresponding values for bulk PST with similar B-site cation order[25] ($T_C \approx 298$ K and $|Q_0| \approx 10.0$ MJ m^{-3}, assuming our density of 8,750 g cm^{-3}). Dielectric measurements (Supplementary Fig. 3c) recorded a Curie temperature of $T_C \approx 292$ K, a small loss tangent of 0.05 just below T_C, and a much smaller loss tangent of $<10^{-3}$ above T_C, where we will now describe large EC effects.

We will first use the well-known indirect method[2] to evaluate EC effects for the active volume (Fig. 2), implementing three improvements (see Methods) that some of us introduced in ref. [19]. Isothermal electrical polarization $P(E)$ was measured every 0.35 K at 385 values of increasing temperature T, and field-removal branches obtained under unipolar conditions ($0 \le E \le 15.8$ V μm^{-1}) were used to plot $P(T,E)$ (Fig. 2a). Inflection points in these field-removal branches were used to identify the phase boundary $E(T)$ and critical endpoint. Combining the gradient $dE/dT = 0.15$ V μm^{-1} K^{-1} of this boundary near T_C with a field-driven polarization change of $|\Delta P| \approx 24.0$ μC cm^{-2} (Supplementary Note 9) implies, via the Clausius–Clapeyron equation $|dE/dT| = |\Delta S_0|/|\Delta P|$, an isothermal entropy change of $|\Delta S_0| \approx 36$ kJ K^{-1} m^{-3}, in good agreement with the thermally driven entropy change of $|\Delta S_0| \approx 35$ kJ K^{-1} m^{-3} (from $|Q_0| \approx 10.2$ MJ m^{-3} and $T_C \approx 290$ K; inset of Supplementary Fig. 3a).

Our dense $P(T,E)$ data (Fig. 2a) were used to evaluate the reversible isothermal entropy change $\Delta S(T,E) = \int_E^0 (\partial P/\partial T)_{E'} dE' > 0$ (Fig. 2b) for field removal ($E \to 0$) at temperature T, using the Maxwell relation $(\partial S/\partial E)_T = (\partial P/\partial T)_E$. The largest field used in our indirect measurements ($E \approx 15.8$ V μm^{-1}) yields an entropy change whose magnitude reaches a peak of $|\Delta S| \approx 43$ kJ K^{-1} m^{-3} near 300 K, while the corresponding refrigerant capacity is $\int_{T_1}^{T_2} |\Delta S(T)| dT \approx 2.8$ MJ m^{-3} (T_1 and T_2 define the FWHM of $|\Delta S(T)|$ for $E \approx 15.8$ V μm^{-1}). Subtracting the entropy change on field removal $\Delta S(T,E) > 0$ from the zero-field entropy $S'(T,0) = S(T,0) - S(270\,\mathrm{K},0)$ (Supplementary Fig. 3b) yields entropy map $S'(T,E)$ (Fig. 2c; adiabatic contours are white). Following an adiabatic contour identifies the reversible adiabatic temperature change for the active regions, yielding $\Delta T(T_s,E) < 0$ (Fig. 2d) for the removal of field E at starting temperature T_s. For $E \approx 15.8$ V μm^{-1}, the peak temperature change that we identify using the indirect method is $|\Delta T| \approx 4.8$ K near 309 K.

Thermocouple measurements of temperature change $\Delta T_{\#}$ versus time t recorded temperature jumps of $|\Delta T_j| \approx 4.0$ K at the MLC face centre, when driving a four-step Brayton cycle for which $E \approx 15.8$ V μm^{-1} and $T_s \approx 315$ K (left inset, Fig. 3a). This well-known cycle comprised (1) a highly adiabatic field-on temperature jump $\Delta T_j > 0$, (2) a slow isofield return to T_s, (3) a highly adiabatic field-off temperature jump $\Delta T_j < 0$ and (4) a slow zero-field return to T_s. The absence of Joule heating, while dumping EC heat in step (2), is important for applications.

Infrared camera measurements at the MLC face centre (300 μm × 300 μm) recorded a slightly larger value of $|\Delta T_j| \approx 4.3$ K (main panel, Fig. 3a), reflecting improved adiabaticity (owing to faster data acquisition by a factor of 11) and reduced thermal mass (approximately 6-μm-thick black paint with relatively low volumetric heat capacity versus the thermocouple affixed with a <0.05 mm^3 drop of black paint). The inactive layers in the active area render this value 10% smaller than the indirectly measured value of $|\Delta T| \approx 4.8$ K for the active volume (Fig. 2d), yielding $|\Delta T_j| \approx 0.90|\Delta T|$ as roughly expected (see Methods). By contrast, inactive thermal mass outside the active area only diminished the magnitude of the temperature change near the periphery of the active area (right inset, Fig. 3a).

Infrared camera measurements at different starting temperatures recorded similar values of $|\Delta T_j|$ for field application/removal (red and blue data, Fig. 3b). The crossover in the magnitude of these field application/removal values (seen previously for Gd)[29] was also observed when varying field (Supplementary Note 10) and is explained in Supplementary Note 11. For our intermediate field of $E \approx 15.8$ V μm^{-1}, values of $|\Delta T_j|$ (red and blue circles, Fig. 3b) match well with values of $0.90|\Delta T|$ from the

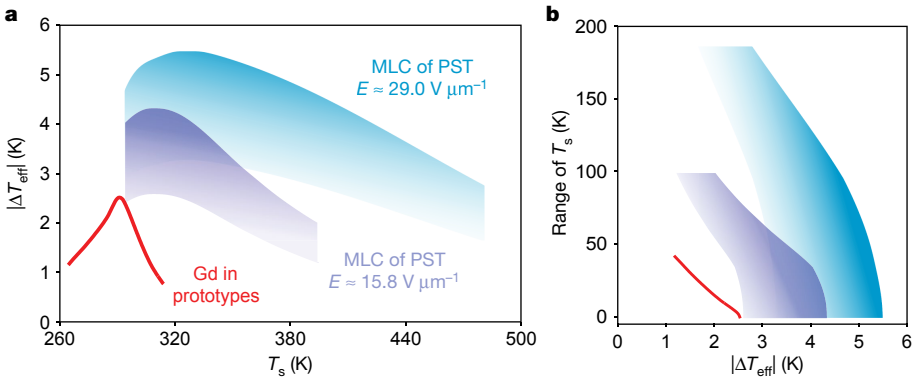


Fig. 4 | Large EC effects over a wide range of operating temperatures. **a**, Effective EC temperature change $|\Delta T_{eff}|$ as a function of starting temperature T_s, for fields of $E = 15.8$ V μm^{-1} and 29.0 V μm^{-1}. Upper bounds from highly adiabatic infrared measurements of $|\Delta T_j|$ (field-off data, Fig. 3b) assume thermalization in the active area alone before useful heat transfer. Lower bounds of $0.60|\Delta T_j|$ assume complete internal thermalization before useful heat transfer. Data are for MLC1. Similar data for five similar MLCs appear in Supplementary Fig. 9. For comparison, we plot $|\Delta T_{eff}|$ versus T_s for a bed of commercial-grade Gd spheres driven with permanent magnets[4], where the internal magnetic field of $\mu_0 H_{int} \approx 1.0$ T corresponds to an applied magnetic field of $\mu_0 H_{app} \approx 1.4$ T (and where μ_0 is the permeability of free space). **b**, The range of T_s for which a given value of $|\Delta T_{eff}|$ is exceeded, with upper and lower bounds shaded as for **a**. This description of peak width provides more information than a single value of refrigerant capacity.

indirect method (black line, Fig. 3b), and slightly exceed values of $|\Delta T_j|$ from thermocouple measurements (brown and black circles, Fig. 3b) due to the improved adiabaticity and reduced thermal mass that we explained earlier. For T_s near 330 K, our maximum field of $E \approx 29.0$ V μm^{-1} yielded our highest value of $|\Delta T_j| \approx 5.5$ K (Fig. 3b). Similar results for five similar MLCs appear in Supplementary Fig. 8.

Our directly measured value of $|\Delta T_j| \approx 5.5$ K exceeds the MC benchmark of $|\Delta T| \approx 2.5$ K for Gd in prototypes (Supplementary Note 1). Moreover, it represents an improvement over other macroscopic EC bodies at any starting temperature, including those for which temperature changes of $|\Delta T_j| \approx 2.2$ K have been directly measured near room temperature (Supplementary Note 2). For the active volume alone, our maximum field of 29.0 V μm^{-1} yields peak values of $|\Delta T| \approx |\Delta T_j|/0.90 \approx 6.1$ K near 330 K, $|\Delta S| = 53$ kJ K^{-1} m^{-3} near 308 K (Supplementary Note 13), and $|Q| = T|\Delta S| \approx 16.6$ MJ m^{-3} near 330 K. The corresponding electrical work $|W| \approx 2.5$ MJ m^{-3} near 330 K (Supplementary Note 14) implies an isothermal materials efficiency $|Q|/|W| \approx 7$ that is similar to the values identified for other EC or elastocaloric materials, and slightly smaller than the values identified for MC materials driven by permanent magnets[30].

Our large EC effects are predicated on the large latent heat associated with high B-site cation order[25,26], such that reducing the B-site cation order reduces the EC effects (Supplementary Note 15), as seen for bulk PST[23]. However, the entropy change ($|\Delta S_0| \approx 35$ kJ K^{-1} m^{-3}) that corresponds to our zero-field latent heat ($|Q_0| \approx 10.2$ MJ m^{-3}) accounts for only about 2/3 of our largest value of $|\Delta S| = 53$ kJ K^{-1} m^{-3}, revealing that there are also substantial EC effects associated with the enhancement of polarization in the transformed phase (see $|\Delta T_j|$ versus E, Supplementary Note 10). Substantial caloric effects associated with a single phase away from absolute zero are well known in mechanocaloric materials (near and away from phase transitions)[31] and in MC materials (near phase transitions)[32], but not in EC materials.

We will now consider the effective temperature change for applications $|\Delta T_{eff}|$, which lies between two bounds that define the shaded regions in Fig. 4a. The upper bound assumes that the active area thermalizes and then exchanges heat with its intended target before exchanging any heat with the inactive area, such that $|\Delta T_{eff}| \approx |\Delta T_j|$. The lower bound assumes that the active volume completely thermalizes with the inactive volume, such that $|\Delta T_{eff}| \approx 0.54|\Delta T| \approx 0.6|\Delta T_j|$ given an active volume of 54% for which $|\Delta T| \approx |\Delta T_j|/0.90$. Upper and lower bounds for the five similar MLCs are presented in Supplementary Fig. 9.

Let us now consider MLC performance with our maximum field ($E \approx 29.0$ V μm^{-1}) while bearing in mind that complete thermalization may be unduly pessimistic because the large, thin active area would make intimate contact with its intended target, for example heat-exchange fluid. Values of $|\Delta T_{eff}|$ peak at 5.5 K if we assume thermalization of the active area only (3.3 K for complete thermalization). Moreover, values of $|\Delta T_{eff}|$ remain large over a wide range of starting temperatures (Fig. 4b), exceeding 3 K for starting temperatures that span 176 K if we assume thermalization of the active area only (73 K for complete thermalization). Large temperature spans have been hitherto accessed through the broad transitions associated with smaller EC effects in relaxors[2], but here we access a large temperature span by driving a first-order transition using supercritical fields (Fig. 2a), as verified for our MLCs by using Landau theory to model the active volume (Supplementary Note 16).

Figure 4 and Table 1 compare $|\Delta T_{eff}|$ with temperature jumps $|\Delta T_{eff}| = |\Delta T_j| = |\Delta T|$ for a bed of commercial-grade Gd spheres[4]. The bed was addressed by fields that could be achieved with permanent magnets[4,5], such that it represents the archetypal MC working body in prototype devices. Our peak upper bound of $|\Delta T_{eff}| \approx 5.5$ K is roughly double the peak value of $|\Delta T_{eff}| \approx 2.5$ K for Gd in prototypes, and our values of $|\Delta T_{eff}|$ fall considerably more slowly when the starting temperature is varied. Even if we conservatively assume both our lower field ($E \approx 15.8$ V μm^{-1}) and complete internal thermalization, our peak value of $|\Delta T_{eff}| \approx 2.6$ K is similar to the peak value for Gd.

In summary, we have demonstrated large EC effects in high-quality MLCs of PST. These effects remain large over a wide range of starting temperatures, such that they could be used for active or passive regeneration in which a large temperature span would be accessed with just a single type of MLC. If MLCs of PST were to replace Gd in MC

Table 1 | Performance summary

Variable	MLC of PST		Gd in prototypes
Driving field	$E = 15.8$ V μm^{-1}	$E = 29.0$ V μm^{-1}	$\mu_0 H_{app} = 1.4$ T
$\lvert\Delta T_{eff}\rvert$	2.6–4.3 K	3.3–5.5 K	2.5 K
$\lvert\Delta S\rvert$	43 kJ K^{-1} m^{-3}	53 kJ K^{-1} m^{-3}	21 kJ K^{-1} m^{-3}

For an MLC of PST driven using electric field E, we present peak upper and lower bounds on effective temperature change $|\Delta T_{eff}|$ (Fig. 4a), and peak entropy change $|\Delta S|$ normalized by active volume (Supplementary Note 13). EC data are for MLC1. For comparison, we present the corresponding data near 291 K for a bed of commercial-grade Gd spheres driven with permanent magnets[4]. The internal magnetic field of $\mu_0 H_{int} \approx 1.0$ T corresponds to an applied magnetic field of $\mu_0 H_{app} \approx 1.4$ T (where μ_0 is the permeability of free space).

prototypes, then the impressive developments of recent decades[1,3] could be exploited by using larger caloric effects over a wider range of temperatures. There would be no need for bulky and expensive permanent magnets[4,5], and the dynamic field profile in an active regenerator[3] could be tailored at will because the constituent MLCs would be individually addressable. Our operating temperatures are relevant for cooling consumer electronics and solar cells, and can be reduced below room temperature by using doped PST[23,33]. MLCs of PST could be further developed by increasing the breakdown field through process optimization, reducing the inactive volume by automated mass production, increasing the active area, and tuning the number of layers to match thermal conductance with system design.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1634-0.

1. Brown, G. V. Magnetic heat pumping near room temperature. *J. Appl. Phys.* **47**, 3673–3680 (1976).
2. Moya, X., Kar-Narayan, S. & Mathur, N. D. Caloric materials near ferroic phase transitions. *Nat. Mater.* **13**, 439–450 (2014).
3. Yu, B., Liu, M., Egolf, P. W. & Kitanovski, A. A review of magnetic refrigerator and heat pump prototypes built before the year 2010. *Int. J. Refrig.* **33**, 1029–1060 (2010).
4. Bjørk, R., Bahl, C. R. H. & Nielsen, K. K. The lifetime cost of a magnetic refrigerator. *Int. J. Refrig.* **63**, 48–62 (2016).
5. Tura, A. & Rowe, A. Concentric Halbach cylinder magnetic refrigerator cost optimization. *Int. J. Refrig.* **37**, 106–116 (2014).
6. Sinyavsky, Y. V., Pashkov, N. D., Gorovoy, Y. M., Lugansky, G. E. & Shebanov, L. The optical ferroelectric ceramic as working body for electrocaloric refrigeration. *Ferroelectrics* **90**, 213–217 (1989).
7. Sinyavsky, Y. & Brodyansky, V. M. Experimental testing of electrocaloric cooling with transparent ferroelectric ceramic as working body. *Ferroelectrics* **131**, 321–325 (1992).
8. Sinyavskii, Y. V. Electrocaloric refrigerators: a promising alternative to current low-temperature apparatus. *Chem. Petrol. Eng.* **31**, 295–306 (1995).
9. Gu, H. et al. A chip scale electrocaloric effect based cooling device. *Appl. Phys. Lett.* **102**, 122904 (2013).
10. Plaznik, U. et al. Bulk relaxor ferroelectric ceramics as a working body for an electrocaloric cooling device. *Appl. Phys. Lett.* **106**, 043903 (2015).
11. Wang, Y. D. et al. A heat-switch-based electrocaloric cooler. *Appl. Phys. Lett.* **107**, 134103 (2015).
12. Sette, D. et al. Electrocaloric cooler combining ceramic multi-layer capacitors and fluid. *APL Mater.* **4**, 091101 (2016).
13. Blumenthal, P., Molin, C., Gebhardt, S. & Raatz, A. Active electrocaloric demonstrator for direct comparison of PMN-PT bulk and multilayer samples. *Ferroelectrics* **497**, 1–8 (2016).
14. Zhang, T., Qian, X.-S., Gu, H., Hou, Y. & Zhang, Q. M. An electrocaloric refrigerator with direct solid to solid regeneration. *Appl. Phys. Lett.* **110**, 243503 (2017).
15. Ma, R. et al. Highly efficient electrocaloric cooling with electrostatic actuation. *Science* **357**, 1130–1134 (2017).
16. Defay, E. et al. Enhanced electrocaloric efficiency via energy recovery. *Nat. Commun.* **9**, 1827 (2018).
17. Qian, S. et al. Performance enhancement of a compressive thermoelastic cooling system using multi-objective optimization and novel designs. *Int. J. Refrig.* **57**, 62–76 (2015).
18. Lloveras, P. et al. Colossal barocaloric effects near room temperature in plastic crystals of neopentylglycol. *Nat. Commun.* **10**, 1803 (2019).
19. Crossley, S., Nair, B., Whatmore, R. W., Moya, X. & Mathur, N. D. Electrocaloric cooling cycles in lead scandium tantalate with true regeneration via field variation. *Phys. Rev. X* **9**, 041002 (2019).
20. Crossley, S. et al. Direct electrocaloric measurement of $0.9Pb(Mg_{1/3}Nb_{2/3})O_3$–$0.1PbTiO_3$ films using scanning thermal microscopy. *Appl. Phys. Lett.* **108**, 032902 (2016).
21. Kar-Narayan, S. & Mathur, N. D. Predicted cooling powers for multilayer capacitors based on various electrocaloric and electrode materials. *Appl. Phys. Lett.* **95**, 242903 (2009).
22. Usui, T. et al. Effect of inactive volume on thermocouple measurements of electrocaloric temperature change in multilayer capacitors of $0.9Pb(Mg_{1/3}Nb_{2/3})O_3$–$0.1PbTiO_3$. *J. Phys. D* **50**, 424002 (2017).
23. Shebanovs, L., Sternberg, A., Lawless, W. N. & Borman, K. Isomorphous ion substitutions and order–disorder phenomena in highly electrocaloric lead-scandium tantalate solid solutions. *Ferroelectrics* **184**, 239–242 (1996).
24. Shebanovs, L., Borman, K., Lawless, W. N. & Kalvane, A. Electrocaloric effect in some perovskite ferroelectric ceramics and multilayer capacitors. *Ferroelectrics* **273**, 137–142 (2002).
25. Stenger, C. G. F. & Burggraaf, A. J. Order–disorder reactions in the ferroelectric perovskites $Pb(Sc_{1/2}Nb_{1/2})O_3$ and $Pb(Sc_{1/2}Ta_{1/2})O_3$. *Phys. Status Solidi A* **61**, 653–664 (1980).
26. Setter, N. & Cross, L. E. The contribution of structural disorder to diffuse phase transitions in ferroelectrics. *J. Mater. Sci.* **15**, 2478–2482 (1980).
27. Imry, Y. & Wortis, M. Influence of quenched impurities on first-order phase transitions. *Phys. Rev. B* **19**, 3580–3585 (1979).
28. Hirose, S. et al. Progress on electrocaloric multilayer ceramic capacitor development. *APL Mater.* **4**, 064105 (2016).
29. Bahl, C. R. H. & Nielsen, K. K. The effect of demagnetization on the magnetocaloric properties of gadolinium. *J. Appl. Phys.* **105**, 013916 (2009).
30. Moya, X., Defay, E., Heine, V. & Mathur, N. D. Too cool to work. *Nat. Phys.* **11**, 202–205 (2015).
31. Lloveras, P. et al. Giant barocaloric effects at low pressure in ferrielectric ammonium sulphate. *Nat. Commun.* **6**, 8801 (2015).
32. Liu, J., Gottschall, T., Skokov, K. P., Moore, J. D. & Gutfleisch, O. Giant magnetocaloric effect driven by structural transitions. *Nat. Mater.* **11**, 620–626 (2012).
33. Whatmore, R. W. et al. Modified lead scandium tantalate for uncooled LWIR detection and thermal imaging. *Proc. 8th IEEE International Symposium on Applications of Ferroelectrics (ISAF)*, 202–205 (IEEE, 1992).

Methods

Samples

All data were obtained using MLC1, except when verifying reproducibility (MLCs 2–6, Supplementary Note 12), evaluating electrical work (MLC7, Supplementary Note 14) and investigating the effect of reduced B-site cation order (MLC8, Supplementary Note 15). The X-ray diffraction and heat capacity data were obtained after MLCs had been crushed to form a powder. The optical image in Fig. 1a was obtained using an MLC that was similar to MLC1.

MLC fabrication

MLCs of PST were prepared by solid-state reaction and tape casting. Stoichiometrically weighed powders of Pb_3O_4, Sc_2O_3 and Ta_2O_5 were ball-milled in distilled water for 17 h with balls of partially stabilized zirconia. The resulting slurry was dried and calcined at 850 °C for 4 h to obtain PST powder. This powder was ball-milled for 24 h in an organic solvent with a binder, and the resulting slurry was used to form green sheets of PST with a 300-μm-gap doctor blade. After screen-printing inner electrodes of Pt paste, we stacked, pressed and cut the green sheets to obtain green chips. Next, the binder was burned off at 500 °C for 4 h. Proto-MLCs were then sintered at 1,400 °C for 4 h and subsequently annealed at 1,000 °C for 100–1,000 h, while surrounded with a mixture of Pb_3O_4 and ZrO_2 powders whose Pb:Zr ratio was 1:1 (the resulting $PbZrO_3$ reaction product prevented lead loss). However, MLC8 with reduced B-site cation order (Supplementary Note 15) did not undergo the anneal at 1,000 °C. Outer electrodes were fabricated with silver paste. The challenges of MLC fabrication are discussed in Supplementary Note 3.

PST grain size and density

An average grain size of 2–3 μm was identified by scanning electron microscopy. We calculated an average value of density $\rho = 8{,}750\ \mathrm{kg\ m^{-3}}$ by using Archimedes' principle to identify $\rho = (w\rho_{water})/(w - v)$ for three samples (5 mm × 5 mm × 0.3 mm) fabricated from thick green sheets without electrodes. Here, w denotes the sample weight in air, v denotes the apparent sample weight in purified water, and ρ_{water} is the density of water.

Temperature control for isothermal measurements of dielectric constant and electrical polarization

We used a bespoke cryogenic probe[34] with which each MLC made good thermal contact. To establish an appropriate timescale for isothermal measurements, the thermal relaxation time (approximately 2.2 s for 1/e decay) was identified at room temperature from infrared measurements of quasi-adiabatic temperature change in air (not shown). The use of vacuum had nominally no effect on thermal relaxation time.

Temperature control for highly adiabatic EC measurements with the thermocouple and infrared camera

The MLC was suspended approximately 0.9 mm above the heater block of a bespoke heating stage[34] that was open to air (Supplementary Fig. 1). The thermal relaxation time (about 7 s for 1/e decay) was identified from thermocouple measurements of temperature change (left inset of Fig. 3a).

Dielectric measurements

These were performed with an Agilent 4294A analyser, by sweeping the frequency from 100 Hz to 100 kHz. Data were collected approximately every 0.12 K while cooling and subsequently heating at the slow rate of $\pm 1\ \mathrm{K\ min^{-1}}$ in our cryogenic probe[34].

X-ray diffraction measurements

These were performed with Cu-Kα radiation using a Bruker D8 Advance diffractometer equipped with a LYNXEYE EX detector. The intensities of the 111 and 200 reflections were determined by fitting a pseudo-Voigt function using HighScore Plus software.

Differential scanning calorimetry

This was performed using a TA Instruments Q2000 DSC that was calibrated via the melting transition of an indium reference sample. Data were obtained on increasing the temperature at 10 K $\mathrm{min^{-1}}$. Following standard practice[35], heat flux $\mathrm{d}Q/\mathrm{d}t$ was normalized by temperature ramp rate $\mathrm{d}T/\mathrm{d}t$ to yield $\mathrm{d}Q/\mathrm{d}T$. Integration of $\mathrm{d}Q/\mathrm{d}T$ yielded latent heat $|Q_0| \approx 10.2\ \mathrm{MJ\ m^{-3}}$ after subtracting a sigmoidal baseline (inset, Supplementary Fig. 3a). Calibration with a sapphire reference permitted $\mathrm{d}Q/\mathrm{d}T$ to be recast as zero-field heat capacity $c(T)$ (Supplementary Fig. 3a). From $c(T)$, we obtained the zero-field entropy $S'(T) = S(T) - S(270\ \mathrm{K}) = \int_{270\ \mathrm{K}}^{T} c(T')/T'\,\mathrm{d}T'$ (Supplementary Fig. 3b) with respect to the entropy S at 270 K.

Indirect EC measurements

Using a Keithley 2410 SourceMeter, highly isothermal measurements of electrical polarization were obtained at constant current (10 μA above T_C, 5 μA below T_C) on warming from 270 K to 405 K in our cryogenic probe[34], such that approximately every 0.35 K we measured one bipolar cycle in ±600 V, two unipolar cycles out to +600 V, and two unipolar cycles out to −600 V. Bipolar cycles were used to centre unipolar cycles on the polarization axis, and $P(T,E)$ data for indirect EC measurements were harvested from the second positive unipolar cycle (examples of bipolar and unipolar cycles appear in Supplementary Fig. 4). The small constant current limited the instantaneous speed of the field-driven transition to yield approximately isothermal conditions (the duration of ≥ 20 s for each cycle branch was ≥9 times as large as the thermal relaxation time of approximately 2.2 s for 1/e decay).

Improvements to the indirect method

Three improvements that some of us introduced elsewhere[19] permit excellent agreement between our indirect EC measurements and our direct measurements (comparisons appear in Fig. 3b and Supplementary Fig. 7). First, $P(E)$ was measured approximately every 0.35 K, in contrast with the standard practice of using 10 K increments, resulting in a dense map of $P(T,E)$ (Fig. 2a) that we used to construct a dense map of $\Delta S(T,E)$ (Fig. 2b). Second, we used unipolar not bipolar $P(E)$ measurements (Supplementary Note 7), thus minimizing field hysteresis to strengthen the single-valued assumption on $P(T,E)$. Third, we evaluated $\Delta T(T_s,E)$ (Fig. 2d) by following adiabatic contours on the entropy map $S'(T,E)$ (Fig. 2c) that we created by subtracting $\Delta S(T,E)$ (Fig. 2b) from $S'(T,0)$ (Supplementary Fig. 3b), thus improving on the standard practice of identifying $\Delta T \approx -T\Delta S/c$ under the assumption of some single effective value for the specific heat capacity c.

Direct thermocouple measurements of temperature change

The measurement set-up appears in Supplementary Fig. 1a. A bespoke K-type thermocouple was monitored at approximately 4.5 Hz to record EC cycles roughly every 5 K on warming from 295 K to 388 K. These cycles were driven using a current of magnitude 10 mA from a Keithley 2410 SourceMeter. The insulation of the 40-μm-diameter thermocouple wires was removed within about 2 cm of the weld to reduce thermal mass. The weld was pressed onto the centre of the MLC face and attached with a drop of matt black paint (PNM400, Electrolube) for good thermal contact. To prepare the MLC for subsequent infrared measurements, the thermocouple and black paint were removed from the MLC by using acetone and then isopropyl alcohol.

Direct infrared measurements of temperature change

The measurement set-up appears in Supplementary Fig. 1b. We used an infrared camera (SC7500, FLIR) operating at 50 Hz to image EC

cycles approximately every 5–10 K on warming in air. These cycles were driven using a current of magnitude 1–10 mA from a Keithley 2410 SourceMeter. To increase emissivity, two layers of matt black paint (PNM400, Electrolube) were spin-coated on the MLC face, resulting in a total thickness of approximately 6 μm. All infrared data represent an average within the 300 μm × 300 μm face centre (white square, Fig. 1b), with the exception of the infrared image that we present in the right inset of Fig. 3a. Supplementary Note 17 explains how we achieved good calibration across starting temperatures that spanned 188 K, and how we identified an emissivity of 0.84–0.87 for the black paint.

Inactive thermal mass in the active area

Inside and not near the periphery of the active area, complete thermalization between the 19 active layers of PST and the inactive layers (two outer layers of PST, 20 inner electrodes of Pt) implies $|\Delta T_j| \approx 0.86|\Delta T|$, assuming an off-peak specific heat capacity of $c \approx 2.7\ \mathrm{MJ\,K^{-1}\,m^{-3}}$ for PST (Supplementary Fig. 3a), and $c \approx 2.8\ \mathrm{MJ\,K^{-1}\,m^{-3}}$ for Pt. The prediction of $|\Delta T_j| \approx 0.86|\Delta T|$ is similar to our empirical finding of $|\Delta T_j| \approx 0.90|\Delta T|$.

Data availability

Source data for Figs. 2–4 are provided with the paper. All other relevant data are available within the paper and its Supplementary Information files.

34. Crossley, S. *Electrocaloric Materials and Devices*. PhD thesis, Univ. Cambridge (2013); http://www.repository.cam.ac.uk/handle/1810/245063.
35. Höhne, G. W. H., Hemminger, W. F. & Flammersheim, H.-J. *Differential Scanning Calorimetry*, 121–126 (Springer, 2003).

Acknowledgements We thank C. Minami, Y. Kojima, N. Furusawa, K. Yamamoto, Y. Inoue and K. Honda for their assistance in fabricating MLCs, and we thank R. Whatmore, À. Torelló and E. Defay for discussions. B.N. is grateful for support from Gates Cambridge, the Winton Programme for the Physics of Sustainability, and Trinity College Cambridge. X.M. is grateful for support from UK EPSRC grant EP/M003752/1, ERC starting grant no. 680032, and the Royal Society. G.G.G.-V. is grateful for support from the Vice-Rectory for Research (project no. B9194) and the Office of International Affairs at the University of Costa Rica, and Churchill College at the University of Cambridge. X.M. and G.G.G.-V. are grateful for support from the Royal Society International Exchanges programme (IES\R3\170025).

Author contributions N.D.M. and X.M. conceived the study and led the project together with B.N. and S.H. T.U. and S.H. were responsible for the fabrication and optimization of the high-quality MLCs. S.C. constructed and commissioned the bespoke apparatus used for electrical and thermal measurements. B.N. performed all of the measurements, except that S.K. performed the X-ray diffraction and obtained the optical images. G.G.G.-V. performed the Landau theory. N.D.M wrote the manuscript and supplementary file, with input and feedback from B.N., X.M., S.H., S.C. and G.G.G.-V.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1634-0.

Correspondence and requests for materials should be addressed to X.M., S.H. or N.D.M.

Peer review information *Nature* thanks Brahim Dkhil, Bai-Xiang Xu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Article

Skin-integrated wireless haptic interfaces for virtual and augmented reality

https://doi.org/10.1038/s41586-019-1687-0

Received: 9 January 2019

Accepted: 2 September 2019

Published online: 20 November 2019

Xinge Yu[1,23], Zhaoqian Xie[1,2,3,4,5,23], Yang Yu[6,7,8,23], Jungyup Lee[9,23], Abraham Vazquez-Guardado[10], Haiwen Luan[10], Jasper Ruban[9], Xin Ning[11], Aadeel Akhtar[12], Dengfeng Li[1], Bowen Ji[3,4,5,13], Yiming Liu[1], Rujie Sun[14], Jingyue Cao[15], Qingze Huo[3,4,5], Yishan Zhong[6,7], ChanMi Lee[6,7], SeungYeop Kim[6,7], Philipp Gutruf[16], Changxing Zhang[17], Yeguang Xue[3,4,5], Qinglei Guo[18], Aditya Chempakasseril[6,7], Peilin Tian[6,7], Wei Lu[10], JiYoon Jeong[9], YongJoon Yu[9], Jesse Cornman[12], CheeSim Tan[12], BongHoon Kim[6,7,10], KunHyuk Lee[10], Xue Feng[17], Yonggang Huang[3,4,5,10]* & John A. Rogers[4,5,6,7,10,19,20,21,22]*

Traditional technologies for virtual reality (VR) and augmented reality (AR) create human experiences through visual and auditory stimuli that replicate sensations associated with the physical world. The most widespread VR and AR systems use head-mounted displays, accelerometers and loudspeakers as the basis for three-dimensional, computer-generated environments that can exist in isolation or as overlays on actual scenery. In comparison to the eyes and the ears, the skin is a relatively underexplored sensory interface for VR and AR technology that could, nevertheless, greatly enhance experiences at a qualitative level, with direct relevance in areas such as communications, entertainment and medicine[1,2]. Here we present a wireless, battery-free platform of electronic systems and haptic (that is, touch-based) interfaces capable of softly laminating onto the curved surfaces of the skin to communicate information via spatio-temporally programmable patterns of localized mechanical vibrations. We describe the materials, device structures, power delivery strategies and communication schemes that serve as the foundations for such platforms. The resulting technology creates many opportunities for use where the skin provides an electronically programmable communication and sensory input channel to the body, as demonstrated through applications in social media and personal engagement, prosthetic control and feedback, and gaming and entertainment.

An important future for VR/AR lies in the development of a full, immersive experience that includes not only interactive images and sounds, but also sensations of touch. The consequences of technologies with multi-sensory capabilities of this type will be far reaching, across fields ranging from social media and communications, to gaming and entertainment, and to clinical medicine, rehabilitation and recovery[1,2]. The skin is the largest organ of the body, and mechanoreceptors distributed across the skin, within the dermis, form the basis of our physical interactions with the world. Specifically, responses to spatio-temporal patterns of force on the skin transmit to the brain as signals that define a mechanical sense of our surroundings[3–5]. Efforts to integrate electronically programmable interfaces to mechanoreceptors within a comprehensive VR/AR platform are, however, in their infancy compared to those associated with video and audio interfaces. Some approaches rely on collections of wired electrodes pressed against the skin to induce artificial, vibration-like sensations via electrostimulation, known as electrotactile effects[4,6,7]. Variability of the impedance of the skin across the body and between individuals, along with time dependent drifts in this quantity due to changes in the properties of the skin or the electrode surfaces, represent confounding challenges in selecting appropriate combinations of voltages and currents that create desired responses without pain or electrically induced lesions[8]. A promising alternative relies on mechanical forces, in the form of vibratory actuation imparted to the skin by electrical motors or piezoelectric devices, where relays, bulk wires and battery packs couple loosely to the body through textiles, tapes and straps to provide the necessary control

[1]Department of Biomedical Engineering, City University of Hong Kong, Hong Kong, China. [2]Department of Engineering Mechanics, Dalian University of Technology, Dalian, China. [3]Department of Civil and Environmental Engineering, Northwestern University, Evanston, IL, USA. [4]Department of Mechanical Engineering, Northwestern University, Evanston, IL, USA. [5]Department of Materials Science and Engineering, Northwestern University, Evanston, IL, USA. [6]Department of Materials Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [7]Frederick Seitz Materials Research Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [8]School of Materials Science and Engineering, Tsinghua University, Beijing, China. [9]NeuroLux Corporation, Evanston, IL, USA. [10]Simpson Querrey Institute, Center for Bio-Integrated Electronics, Northwestern University, Evanston, IL, USA. [11]Department of Aerospace Engineering, Pennsylvania State University, University Park, PA, USA. [12]PSYONIC, Inc., Champaign, IL, USA. [13]Department of Micro/Nano Electronics, Shanghai Jiao Tong University, Shanghai, China. [14]Bristol Composites Institute, University of Bristol, Bristol, UK. [15]Wearifi, Inc., Evanston, IL, USA. [16]Department of Biomedical Engineering, University of Arizona, Tucson, AZ, USA. [17]AML, Department of Engineering Mechanics, Interdisciplinary Research Center for Flexible Electronics Technology, Tsinghua University, Beijing, China. [18]School of Microelectronics, Shandong University, Jinan, China. [19]Department of Biomedical Engineering, Northwestern University, Evanston, IL, USA. [20]Department of Neurological Surgery, Feinberg Medical School, Northwestern University, Evanston, IL, USA. [21]Department of Chemistry, Northwestern University, Evanston, IL, USA. [22]Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA. [23]These authors contributed equally: Xinge Yu, Zhaoqian Xie, Yang Yu, Jungyup Lee. *e-mail: y-huang@northwestern.edu; jrogers@northwestern.edu

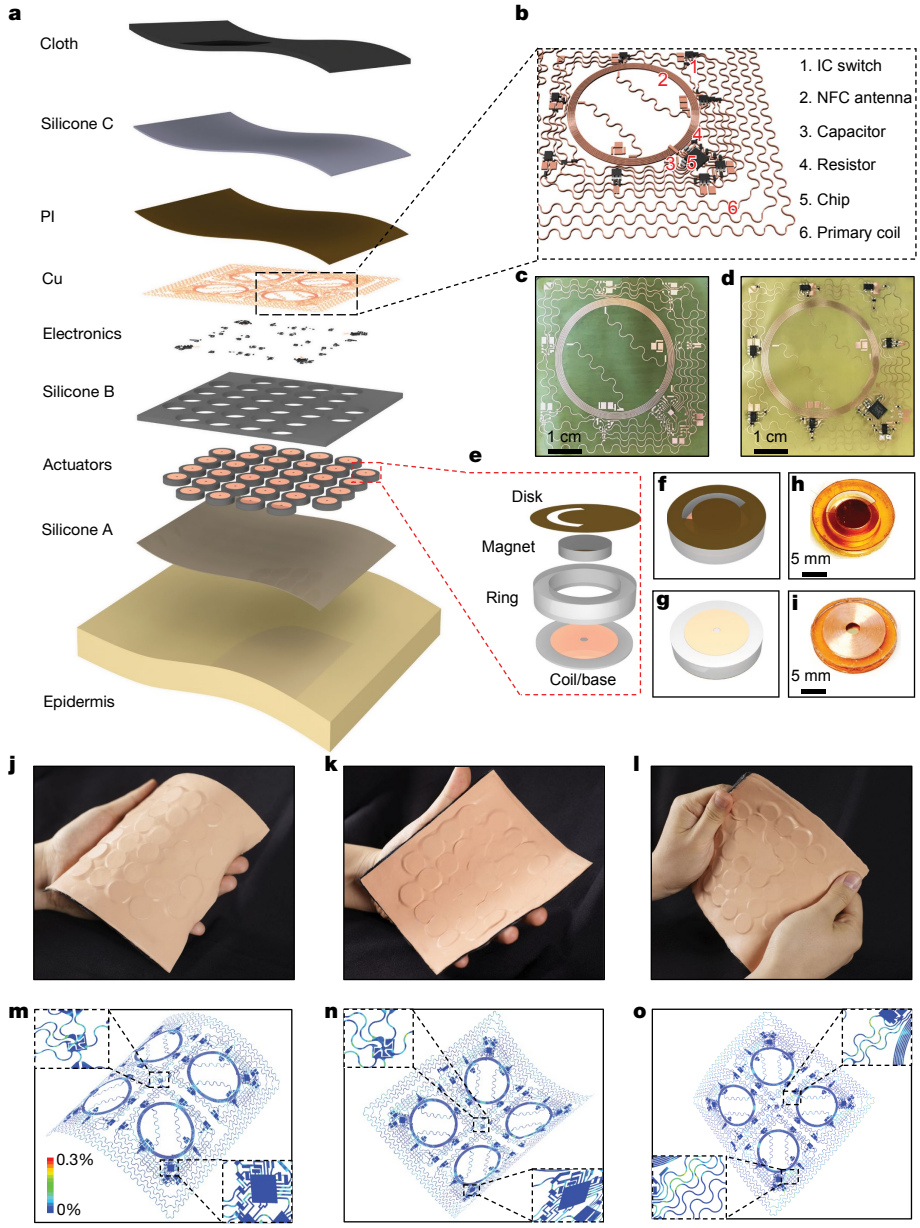


Fig. 1 | Design and architecture of an epidermal VR system. a, Exploded-view schematic illustration of a device with 32 independently controlled haptic actuators. **b**, Schematic illustration of the NFC electronics and circuit; the main circuit components are labelled 1–6. **c**, **d**, Optical images of an NFC coil before (**c**) and after (**d**) integrating the electronic components. **e**, Exploded-view schematic diagram of a haptic actuator. **f**, **g**, Schematic diagram of an actuator viewed from above (**f**) and below (**g**). **h**, **i**, Optical images of an actuator viewed from above (**h**) and below (**i**). **j**–**o**, Optical images (top row) and FEA results (bottom row) of an epidermal VR device under bending (**j**, **m**), folding (**k**, **n**) and twisting (**l**, **o**). The colour in **m**–**o** represents the equivalent strain, and the insets show the areas with relatively high strain levels. See Methods for details.

systems and power supplies[2,7]. As with related supporting hardware for electrotactile interfaces, the cumbersome nature of this type of technology and the limited ability to scale to monolithic, manufacturable platforms with large numbers of independently controlled actuators represent disadvantages that will hinder widespread adoption.

Here we introduce a set of materials, device designs, integration schemes and system layouts for wirelessly controlled and wirelessly powered, battery-free, haptic interfaces that incorporate large arrays of millimetre-scale vibratory actuators in soft, conformal sheets of electronics that laminate directly onto the skin in a simple, non-invasive and reversible manner. Multiple systems of this type, interfaced onto desired locations on the body with full, programmable control via a remote computer system, establish means to extend VR/AR experiences beyond visual and auditory sensations, with broad application possibilities. Figure 1a presents schematic illustrations of a representative platform, which we refer to as an epidermal VR interface. The construction takes the form of a multilayer stack that includes (1) a thin elastomeric layer as a reversible, soft, adhesive interface to the skin, (2) a silicone-encapsulated functional layer that supports a wireless control system, a means for receiving wirelessly transmitted power, and an interconnected array of actuators with associated drive electronics, and (3) a breathable, stretchable fabric coated with an thin film of silicone, as a physically tough but skin-conformal supporting substrate with strain-limiting mechanics to prevent damage to the functional materials and components. For aesthetics, a coating of silicone that incorporates skin-tone colouration and/or graphics can be included on the outward facing side of the fabric, or such features can be incorporated directly into the fabric itself. The electronics part of the functional layer consists of a collection of copper (Cu) traces encapsulated in polyimide (PI) and formed in narrow, filamentary serpentine geometries

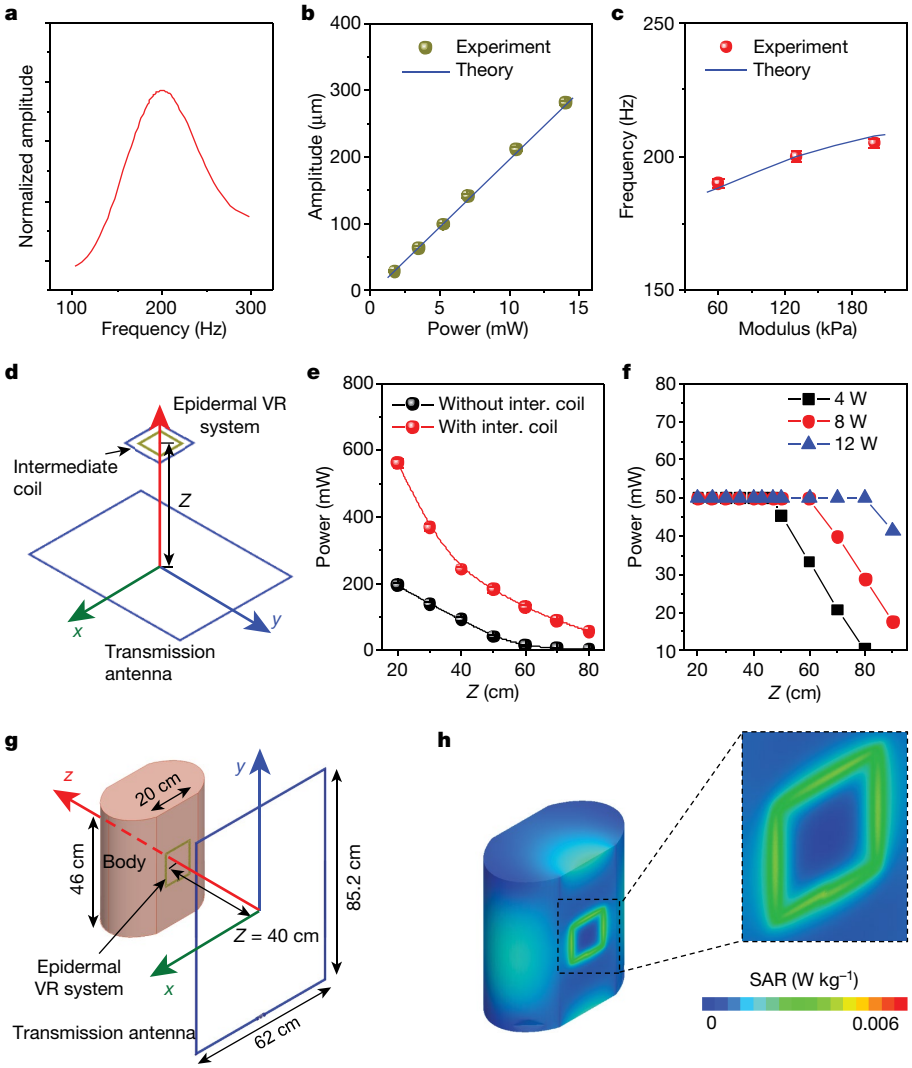


Fig. 2 | Optimized operation of key electrical and mechanical components of an epidermal VR system. a, Normalized amplitude–frequency response of a haptic actuator in contact with a skin phantom. **b**, Travel amplitude of the magnet as a function of the input power (data points) for an actuator in contact with skin phantoms with elastic moduli of 130 kPa. Here, and in **c**, the symbols and lines correspond to experimental and mechanics-simulated theoretical results, respectively. **c**, Dependence of the resonance frequency on the elastic modulus of the skin phantom (data points), over a range relevant for human skin. **d**, Schematic illustration of an epidermal VR device with an intermediate coil above a transmission antenna. **e**, Power harvested from the primary coil of an epidermal VR device with and without an integrated intermediate ('inter.') coil, as a function of distance in the *Z* direction. Here, and in **f**, the lines are guides to the eye. **f**, Power harvested from the primary coil of an epidermal VR device with an intermediate coil after power regulation, for various RF powers applied to the transmission antenna (4–12 W; see key). **g**, **h**, Schematic illustration of a representative epidermal VR system mounted on the body (**g**), and the SAR distribution (**h**). In **b**, **c**, **e**, **f**, error bars correspond to the calculated standard deviation.

according to quantitative design rules in stretchable electronics[9,10]. These traces interconnect a collection of small, chip-scale integrated circuit components and passive elements, including magnetic radio frequency (RF) loop antenna structures, resistors, capacitors, rectifiers and integrated circuit (IC) switches. System-on-a-chip (SoC) ICs that include microcontrollers with capabilities in near-field communication (NFC) and general input/output functionality (Fig. 1b–d) serve as control interfaces to a distributed set of mechanical vibratory actuators, referred to in the following as haptic actuators.

Supplementary Fig. 1e–i shows schematic diagrams and optical images of these actuators. Here, time-dependent Lorenz forces (Extended Data Fig. 1) follow from the passage of a time-varying current through a coil that surrounds a permanent magnet. The shell of the actuator consists of a ring-shaped elastomeric structure that provides space for the magnet to travel freely in the out-of-plane direction. A thin disk of PI mounted on top of the PDMS ring and laser-cut with a semicircular slit serves as a bonding location for the magnet. This construct forms a cantilever-like platform, capable of actuation via interactions between the magnet and current flowing through the coil at the base of the ring. These basic designs can, in principle, be extended to length scales that characterize the separation of mechanoreceptors in the skin of the arms, chest, back and legs (Supplementary Fig. 2). Moreover, miniaturization of this type of actuator by an order of magnitude increases the acceleration of the magnet during vibration by more than a factor of three for the same electrical power, such that the same contact pressure can be achieved by reducing the radial and thickness dimensions of the magnet by a factor of 10 and 3, respectively. As a result, further decreases in power consumption might be achievable by reducing the sizes of the actuators. (See details of scaling simulations in Methods.)

Careful optimization of the materials and designs of these actuators, guided by computational modelling, allows for power efficient operation as skin-coupled haptic interfaces. The diameter and thickness of the PI disk and the layout of the slit, the geometry of the PDMS ring,

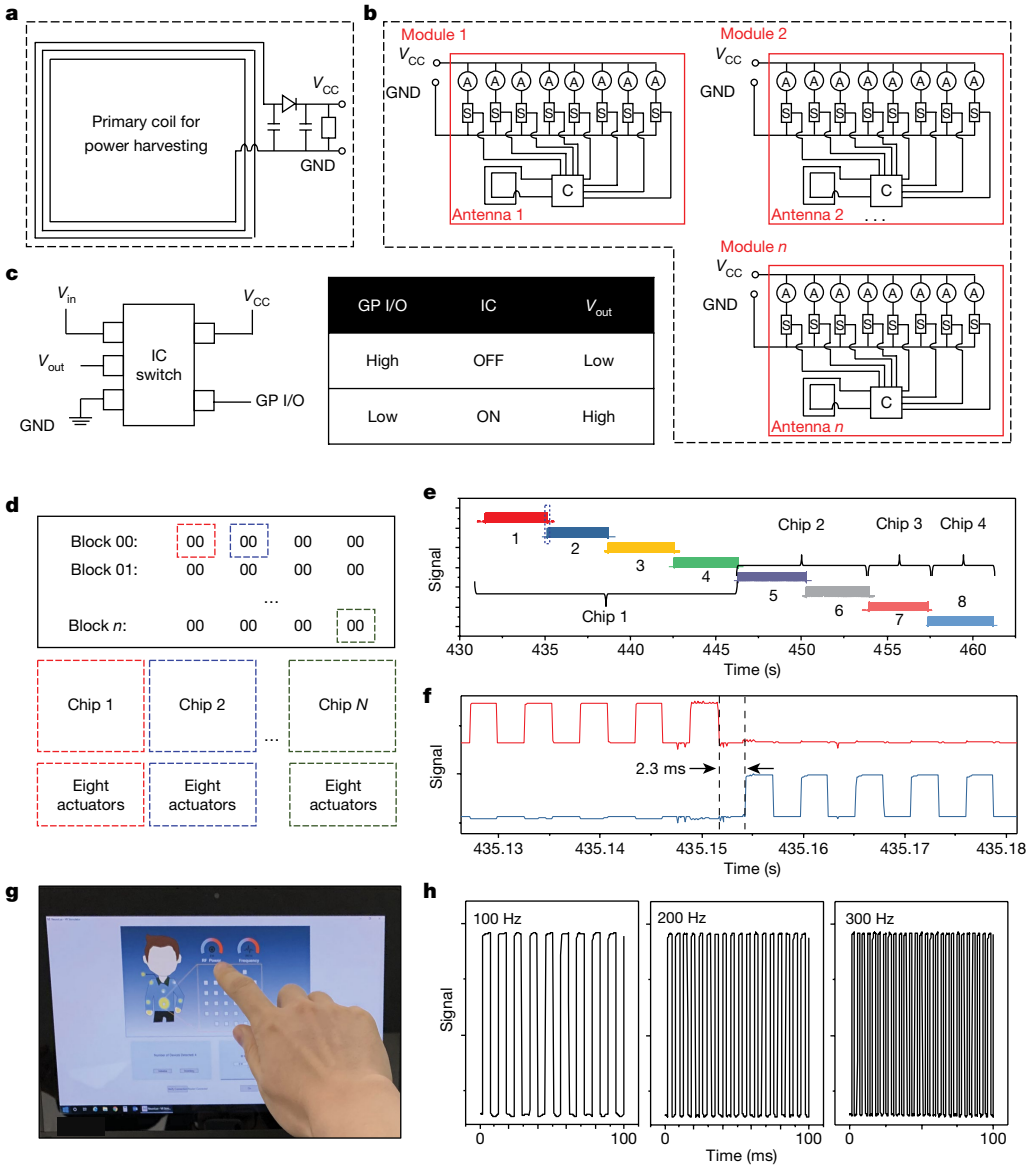


Fig. 3 | Wireless control strategies for epidermal VR systems. a, **b**, Circuit diagram for an epidermal VR device, including a large primary coil for power harvesting (**a**) and several control modules (**b**), each of which consists of a small antenna and an SoC, along with eight haptic actuators controlled by eight IC switches, independently. GND, ground; A, actuator; S, IC switch; C, SoC. **c**, Schematic diagram (left) and working principle (right) of an IC switch. The ON/OFF of the output voltage of the IC switch is controlled by GP I/O ports on the SoC. **d**, Diagram of the command interface that supports independent control over every actuator in the system. Each of the eight GP I/O ports in each SoC is defined by a one-byte command, such that $8 \times n$ actuators can be initiated in any form by a portfolio of n bytes. The dashed coloured boxes represent different SoCs. **e**, Response time of actuators controlled by four SoCs. The coloured plots represent different output signals of the actuators. **f**, A magnified view of the time required to switch from one actuator to another. **g**, Software interface of the control system. **h**, Three representative working frequencies for haptic actuators in an epidermal VR platform: 100 Hz, 200 Hz and 300 Hz.

the type and size of the magnet, and the configuration of the coil can be selected to satisfy requirements relevant to body-interfaced operation, across many mounting locations and body types, as outlined in quantitative detail in Extended Data Figs. 1–4. Frequencies in the range between 100 Hz and 300 Hz are of greatest interest because they provide the strongest sensations on human skin, owing to the intrinsic nature of the responses of the mechanoreceptors[11,12]. Here, amplitudes as small as several micrometres can yield distinct tactile responses[11]. Adjusting the angular extent of the slit in the PI disk presents a simple means to tune the resonant frequency of the actuator (Extended Data Figs. 2, 3) to a value of 200 Hz for operation on a skin phantom with an elastic modulus of 130 kPa. With optimized designs, these types of haptic actuators require only about 1.75 mW to induce a notable sensory responses on the fingertips and hands, with a corresponding amplitude of approximately 35 μm (Supplementary Videos 1, 2; Fig. 2b), without parasitic heating effects (Supplementary Fig. 3) and with the option of full system operation with small batteries. By comparison, widely used commercially available vibration actuators/motors (eccentric rotating mass actuators, linear resonant actuators and piezoelectric actuators) in consumer gadgets typically require >100 mW. Figure 2c and Extended Data Figs. 3, 4 summarize the dependence of the resonance frequency on the modulus of the phantom, from 60 kPa to 200 kPa. The results suggest a weak, almost negligible, variation in frequency over modulus values that span those characteristic of skin at different ages and across different regions of the body. The amplitude increases linearly with input power for all modulus values (Extended Data Figs. 3e, 4e). The motion of the actuator involves a vibrational deflection along the cantilever beam in a way that directly, and indirectly through inertial

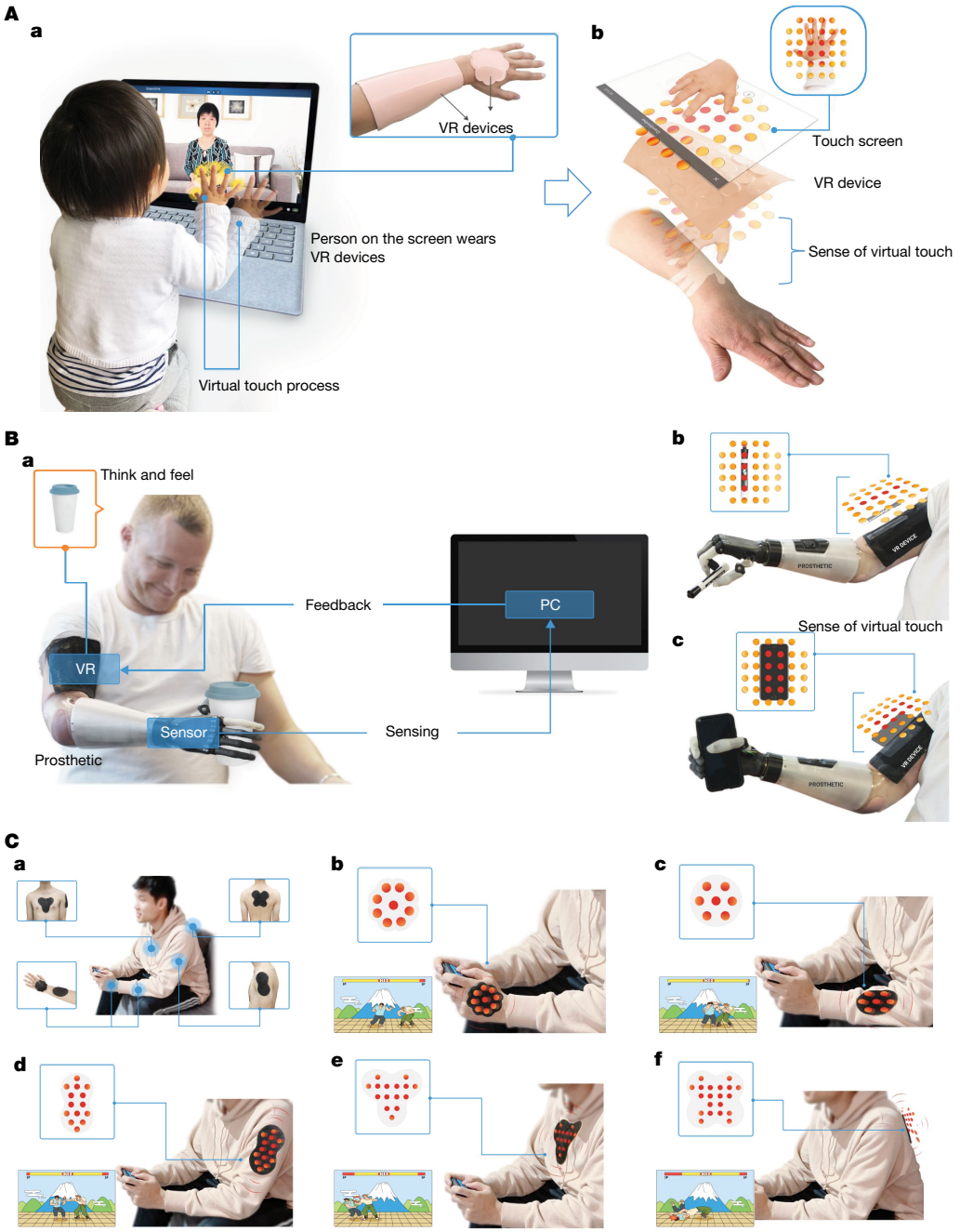


Fig. 4 | Examples of applications of epidermal VR systems. A, Social media application: **a**, a girl touches a screen that displays a video feed of her grandmother, who is wearing an epidermal VR device on her hand and her arm (inset photograph); **b**, a dynamic illustration of the pattern of 'virtual touch process' and 'sense of virtual touch'. **B**, Prosthetics application: **a**, a man with a lower-arm amputation wears a prosthetic arm with a robotic hand and an epidermal VR device on his upper arm; **b**, **c**, the device produces a haptic pattern of sensation ('think and feel') that reproduces the shape characteristics of objects ('feedback') held in the robotic hand ('sensing'). **C**, Gaming application: **a**, a man wears several epidermal VR devices on different parts of his body; **b**–**f**, devices activate when a strike occurs on the corresponding body part of the game character, namely, the hand (**b**), elbow (**c**), arm (**d**), chest (**e**) and back (**f**).

effects, couples to the skin. Careful optimization minimizes mechanical cross-talk between adjacent actuators (Extended Data Figs. 4f, 5).

These actuators connect to associated antenna structures and electronic components through conductive traces with designs that minimize strains and resistive losses and, at the same time, operate without failure under a full range of bending and twisting motions. As with actuator design, 3D finite-element analysis (FEA) techniques guide selection of the interconnect geometries and the overall system layouts (see Methods for details). The result is a soft, deformable platform capable of establishing a comfortable, non-irritating interface to the skin, across nearly any region of the body[13–15]. Computed distributions of strain in the copper and corresponding optical images in Fig. 1j–o show results for bending, folding and twisting. The equivalent strains remain below the elastic limit (0.3%) for a bend of about 145° (with a bending radius of approximately 5.1 cm), a fold of about 150° (folding radius approximately 5 cm) and a twist of about 50° (ref. [16]). (Fig. 1l shows a 'bend', Fig. 1k shows a 'fold'.) The overall shapes of the platforms can be designed to harmonize with anatomical features; examples include 'flower', 'oval', 'peanut', 'triangle' and 'butterfly' shapes (Supplementary Fig. 2).

NFC protocols serve as the basis for operation and coordinated control, with modes that are difficult or impossible to reproduce with battery-free far-field techniques[17]. Power delivery and data communication use different antennas, designed to allow independent operation without interference (Extended Data Fig. 6a, b)[18]. A large primary coil that exploits the full perimeter of the device platform harvests the power needed to operate the entire collection of actuators. Separate small antennas serve as wireless interfaces and power sources for each of the SoCs, each one of which supports independent control over eight actuators. Means for efficient harvesting and utilization of power are critical aspects of system design[18,19]. Most envisioned applications require wireless power transfer over distances of at least tens of centimetres, to allow integration of the RF power transmission antenna into the base of a chair, desk or bed[13]. As summarized in Extended Data Fig. 6c–i, the operating distance (Z) of a VR device (using a 12 cm × 12 cm square serpentine coil as the primary coil with a Q factor of 20, and 32 actuators all working at a set power of 1.75 mW) oriented parallel to the antenna is about 30 cm and 45 cm for transmission antennas with respective dimensions of 31.8 cm × 33.8 cm and 85.2 cm × 62.0 cm, and an input power of 12 W to the transmission antenna. The distance Z can be increased by increasing the power, optimizing the size of the transmission antenna and/or by decreasing the power consumption of the system by reducing the sizes of the actuators.

Addition of an intermediate, single-loop coil (20 × 20 cm; tuned to 13.56 MHz) incorporated directly into the device (Fig. 2d, Supplementary Fig. 4) further improves this range by locally increasing the strength of the magnetic field by a factor of about 10 (Extended Data Fig. 7). Figure 2d compares the power harvested by the primary coil as a function of position Z for cases with and without the intermediate coil, using a transmission antenna (85.2 cm × 62.0 cm) powered at 12 W. The intermediate coil increases the magnetic field strength by about 16 times for Z = 30 cm, and by about 11 times for Z = 50 cm. The corresponding increases in received power are approximately 2.5 times for Z = 30 cm, and 4.5 times for Z = 50 cm, influenced by the load (resistance) of the device[20]. The working range can reach 80 cm. An active power regulation system enabled by a linear voltage regulator ensures consistent operation throughout this full range of distances. Figure 2f highlights the ability to deliver a fixed output power from the primary coil of the device into a load resistor (2 kΩ) for operation of the transmission antenna at powers between 4 W and 12 W across a range of distances. The power output is stable for tilt angles up to 60° and after more than 10,000 cycles of bending to a radius of 2.8 cm (Extended Data Fig. 8, Supplementary Fig. 5). As shown in Fig. 2g, h and Supplementary Fig. 4, the mode of operation complies with guidelines outlined by the Federal Communications Commission (47 CFR Part 1.1310 and 15) and the Federal Drug Administration in terms of both the specific absorbed radiation (SAR) and the maximum permissible exposure (MPE). The maximum value of the SAR is 0.006 W kg^{-1} (Fig. 2h), substantially less than the exposure limit[21] of 0.08 W kg^{-1}. The maximum computed equivalent power density of electromagnetic fields is 0.8 mW cm^{-2} (Supplementary Fig. 6), which is below the MPE limit[21] of approximately 4.9 mW cm^{-2}.

The block diagram in Fig. 3a, b summarizes the system architecture and overall operation. The first part harvests power through a primary coil (Fig. 3a) via the transmission antenna. This power passes through a linear voltage regulator to provide a fixed, direct-current voltage (V_{CC}, where CC indicates common collector) to all of the haptic actuators. The second part provides control and communication via an array of interconnected SoCs, each with a separate small control antenna, and with operational control over eight haptic actuators through its general purpose input/output (GP I/O) ports (Fig. 3b). These ports generate square-wave signals by alternating the output of each GP I/O between its high and low settings, at programmable frequencies between 100 Hz and 300 Hz. An IC switch associated with each actuator (Fig. 3c) transforms V_{CC} into a square wave defined by the corresponding GP I/O, and this signal serves as input to the actuator. In this way, one SoC/IC switch combination allows for control/operation of eight actuators, simultaneously and independently. Scaling this architecture to include multiple unit cells of this type yields systems with arbitrarily large numbers of actuators, without limitation.

Wirelessly writing the necessary NFC Data Exchange Format (NDEF) messages into each SoC via the transmission antenna, which serves simultaneously as an RF reader, defines the output frequency to each haptic actuator, thus programming the entire system for operation. Unlike conventional multiplexing approaches to controlling actuators, a computer interfaced to the reader collects information on the identification codes for each of the SoCs in the system, thereby identifying every GP I/O port and haptic actuator. A one-byte message sets the ON/OFF command for each of the eight GP I/O ports on each SoC. All GP I/O ports are controlled independently. An entire system with 32 haptic actuators (4 SoCs, each with 8 GP I/Os) can be controlled in a single communication of four bytes—for example, a command of 'FF FF FF FF' sets all of the actuators to ON (Fig. 3d), and a command '01 00 00 00' sets to ON the actuator connected to the first GP I/O of the first SoC.

The time required to change from one system configuration to another is in the millisecond regime, as shown in Fig. 3e, f, which is about 50 times faster than the reaction time to tactile stimulation[22]. A graphical interface with a touch screen, as in Fig. 3g and Supplementary Fig. 7, allows a user to change patterns of actuation rapidly and, separately, to select the amplitudes and frequencies of the vibratory responses. An array of 32 lasers provides a means to visualize operation at the system level, via projection of individual beams reflected from each of the haptic actuators and onto a monitoring screen (Extended Data Figs. 9, 10 and Supplementary Video 3). Recordings from a high-speed camera replayed at 16.7× and 133× slow motion reveal the time dynamics of the vibratory motions of the actuators and their programmed control (Supplementary Video 4).

The transmission antennas and devices can be configured in various ways for different use scenarios (Supplementary Fig. 8). Replacing the IC switches with high-power compliance transistors and modifying the cantilever designs improves the power delivery to the actuators and increases their vibratory bandwidths and amplitudes, for enhanced sensation across different body types and anatomical locations. A single-stage voltage regulation scheme conditions the power-harvesting efficiencies of receiving antennas with different form factors. Figure 4A summarizes a possibility in virtual interactions via social media. Here, a girl virtually touches her grandmother's hand through an interface on the screen of a laptop that simultaneously displays a video/audio feed. In this example, two epidermal VR devices are mounted on the grandmother, who experiences a haptic sensation in the form of a continuous wave of vibratory excitation extending sequentially down from her forearm to her hand in a spatio-temporal pattern of touch to match that of the granddaughter's fingertips on the image on the touch screen. A second representative application is in tactile feedback for use of robotic prosthetic devices. Figure 4B shows a man, whose lower arm has been amputated, with an epidermal VR device on his residual limb as he uses a prosthetic arm to grasp objects. Here, sensors on the prosthetic detect the shape of the object and this information serves as input to create a virtual haptic representation of the shape on his upper arm. The third application is in haptic engagement in gaming. In the example of Fig. 4C, a gamer wears several epidermal VR devices across different locations of the body. As a strike occurs in this combat game, haptic actuation reproduces the pattern of the impact at a corresponding location.

The epidermal VR systems introduced here exploit thin, soft architectures capable of laminating directly onto the skin as a platform for programmable control of large arrays of miniaturized haptic actuators in wireless modes of operation and with lightweight, battery-free designs. This class of technology is qualitatively distinguished in form and function over previous attempts at programmable haptic interfaces

to the body. Comprehensive experimental and computational studies of the various subsystems in these platforms yield a basic understanding of their operation and a set of guidelines for design choices. Demonstrations in social media interactions, prosthetic feedback and video gaming are representative of a broad spectrum of potential applications, which also include systems for personalized rehabilitation, surgical training, educational feedback, and multimedia entertainment experiences. Many opportunities exist to improve the performance of these systems by increasing the strength of mechanical actuation at the skin interface.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1687-0.

1. Kapur, P., Jensen, M., Buxbaum, L. J., Jax, S. A. & Kuchenbecker, K. J. Spatially distributed tactile feedback for kinesthetic motion guidance. In *2010 IEEE Haptics Symp*. (ed. Basodogan, C.) 519–526 (IEEE, 2010).
2. Sardini, E., Serpelloni, M. & Pasqui, V. Wireless wearable T-shirt for posture monitoring during rehabilitation exercises. *IEEE Trans. Instrum. Meas.* **64**, 439–448 (2015).
3. Kim, Y. et al. A bioinspired flexible organic artificial afferent nerve. *Science* **360**, 998–1003 (2018).
4. Novich, S. D. & Eagleman, D. M. Using space and time to encode vibrotactile information: toward an estimate of the skin's achievable throughput. *Exp. Brain Res.* **233**, 2777–2788 (2015).
5. Sundaram, S. et al. Learning the signatures of the human grasp using a scalable tactile glove. *Nature* **569**, 698–702 (2019).
6. Withana, A., Groeger, D. & Steimle, R. Tacttoo: a thin and feel-through tattoo for on-skin tactile output. In *Proc. 31st Ann. ACM Symp. on User Interface Software and Technology* (ed. Wilson, A. D.) 365–378 (ACM, 2018).
7. Huang, C.-T., Shen, C.-L., Tang, C.-F. & Chang, S.-H. A wearable yarn-based piezo-resistive sensor. *Sens. Actuators A* **141**, 396–403 (2008).
8. Cogan, S. F. Neural stimulation and recording electrodes. *Annu. Rev. Biomed. Eng.* **10**, 275–309 (2008).
9. Song, Y. M. et al. Digital cameras with designs inspired by the arthropod eye. *Nature* **497**, 95–99 (2013).
10. Rogers, J. A., Someya, T. & Huang, Y. Materials and mechanics for stretchable electronics. *Science* **327**, 1603–1607 (2010).
11. Wilska, A. On the vibrational sensitivity in different regions of the body surface. *Acta Physiol. Scand.* **31**, 285–289 (1954).
12. Mahns, D. A., Perkins, N. M., Sahai, V., Robinson, L. & Rowe, M. J. Vibrotactile frequency discrimination in human hairy skin. *J. Neurophysiol.* **95**, 1442–1450 (2006).
13. Han, S. et al. Battery-free, wireless sensors for full-body pressure and temperature mapping. *Sci. Transl. Med.* **10**, eaan4950 (2018).
14. Kim, D.-H. et al. Epidermal electronics. *Science* **333**, 838–843 (2011).
15. Kim, J. et al. Battery-free, stretchable optoelectronic systems for wireless optical characterization of the skin. *Sci. Adv.* **2**, e1600418 (2016).
16. Zhang, Y. et al. Experimental and theoretical studies of serpentine microstructures bonded to prestrained elastomers for stretchable electronics. *Adv. Funct. Mater.* **24**, 2028–2037 (2014).
17. Huang, X. et al. Epidermal radio frequency electronics for wireless power transfer. *Microsyst. Nanoeng.* **2**, 16052 (2016).
18. Chung, H. U. et al. Binodal, wireless epidermal electronic systems with in-sensor analytics for neonatal intensive care. *Science* **363**, eaau0780 (2019).
19. Beh, T. C., Kato, M., Imura, T., Oh, S. & Hori, Y. Automated impedance matching system for robust wireless power transfer via magnetic resonance coupling. *IEEE Trans. Ind. Electron.* **60**, 3689–3698 (2013).
20. Kim, J., Son, H., Kim, K. & Park, Y. Efficiency analysis of magnetic resonance wireless power transfer with intermediate resonant coil. *IEEE Antennas Wirel. Propag. Lett.* **10**, 389–392 (2011).
21. The Electronic Code of Federal Regulations (e-CFR) Title 47, 1.1310. Radiofrequency radiation exposure limits. https://www.ecfr.gov/cgi-bin/retrieveECFR?gp=&SID=81a1c5a13f8d02f165a06af36937d612&mc=true&n=pt47.1.1&r=PART&ty=HTML#se47.1.1_11310 (e-CFR, 2019).
22. Akamatsu, M., Mackenzie, I. S. & Hasbroucq, T. A comparison of tactile, auditory, and visual feedback in a pointing task using a mouse-type device. *Ergonomics* **38**, 816–827 (1995).

Methods

Fabrication of the system of electronics

A sheet of polyimide (PI, 12.5 μm) coated with a thin layer of copper (Cu, 50–200 μm wide and 18 μm thick) served as the substrate for the antenna structures and the electrical interconnects. Photolithography and etching yielded patterns of Cu in the desired geometries. For systems smaller than 4 inches × 4 inches, the process used a positive photoresist (AZ P4620, AZ Electronic Materials) spin-cast at 3,000 r.p.m. for 30 s, soft baked on a hot plate at 110 °C for 4 min, exposed to ultraviolet (UV, wavelength 350–400 nm) light to a dose of 500 mJ cm^{-2}, and developed for ~70 s in a basic solution (AZ 400K/deionized (DI) water in a 1:3 volume ratio). For systems larger than 4 inches × 4 inches, the process used dry film photoresist (Dupont, 38 μm thick), bonded onto the Cu foil by a roll laminator at 110 °C, soft baked on a hot plate at 110 °C for 4 min, exposed to UV light to a dose of 500 mJ cm^{-2}, post baked on a hot plate at 110 °C for 2 min and developed for ~180 s in a basic solution (MIF 917/DI water in a 1:1 volume ratio), and wet etched (CE-100 copper etchant, Transene) for ~2 min with frequent rinsing with DI water. In both cases, the photoresist was removed by acetone and the substrates were then rinsed with DI water.

Fabrication of the haptic actuators

The first step involved placing a Cu coil (wire diameter of 50 μm, with 300 turns to form a coil with an inner diameter of 3 mm and an outer diameter of 14 mm (Yisu Electronics, Inc.)) in the centre of an acrylic mould (mould 1) with a silicone release reagent (Clearco Product Co., Inc.). Pouring a prepolymer to poly(dimethylsiloxane) (PDMS; Sylgard 184, Dow-Corning; 10:1 weight ratio of prepolymer to crosslinker) into mould 1 submerged the Cu coil under a layer of PDMS with modulus of ~1 MPa and thickness of 0.2 mm. Baking in an oven at 70 °C for 1 h cured the material into a solid, elastomeric form. Next, filling with additional PDMS prepolymer, and mounting in a second, matching mould (mould 2) held in place with set screws prepared the assembly for a second curing step (overnight in an oven at 70 °C) to seal the coil structure in PDMS (~1 MPa; inner and outer diameters of 18 mm and 12 mm, respectively; thickness of 2.5 mm), shaped to meet the design requirements. The dimensions of the resulting PDMS ring with coil in its base, formed by release from moulds 1 and 2, were: 2.4 mm in thickness and 18 mm in diameter, with an inner cavity of 2.2 mm depth and 12 mm diameter (Extended Data Fig. 2). The second part of the actuator consisted of a permanent magnet (nickel-plated neodymium magnet, diameter of 8 mm, thickness of 1.6 mm) mounted on a PI disk. Laser cutting formed circular shapes with diameters of 18 mm from sheets of PI (125 μm thick, DuPont) and semicircular slits with diameters of 8 mm and central angles of 217° (Extended Data Fig. 2). A strong double-sided adhesive (Kapton, DuPont) attached a disk-shaped magnet (nickel-coated neodymium magnet, diameter of 8 mm, thickness of 1.6 mm, weight 0.6 g, Bunting Magnetics Co.) on the cantilever part of the PI disk (18 mm diameter, 125 μm thick). The final step involved bonding the PI disk, with the magnet mounted on the back side, on top of the PDMS ring with a silicone adhesive (Kwik-Sil, WPI Inc.). Each completed actuator weighs 1.4 g. The frequency of the current input to the coil defined the frequency of vibration of the magnet. The magnitude of the current was controlled at a set point (~5 mA for experiments reported here).

Scaling simulation of the haptic actuators

Careful selection and optimization of various parameters of the actuators and the entire VR system focused partly on generating a sufficiently strong alternating magnetic field with the Cu coil to excite suitable vibrations of the magnet. According to the magnetic field distribution around the Cu coil[23], the acceleration a of the magnet can be expressed as

$$a \propto \frac{\sqrt{P_{\text{input}} SNn}\, D_{\text{outer}}^{1.5}}{\left(D_{\text{outer}}^2 + d^2\right)^{1.5}}$$

where P_{input} is the input power, S is the cross-sectional area of the copper wire, N is the layer number, n is the turn number of each layer, D_{outer} is the outer diameter of the Cu coil, and d is the distance between the Cu coil and the magnet (Extended Data Fig. 2). Further, the contact pressure p can be expressed as

$$p = a\rho h$$

where ρ and h are the mass density and the thickness of the cylindrical magnet, respectively. The pressure distribution on the skin associated with operation of the actuator can be calculated by FEA, as in Supplementary Fig. 9. These results suggest that the force f applied on the skin for an actuator with input power of 1.75 mW is 135 mN, from $f = \int_s p\,ds$. To examine the effects of miniaturization, we fix P_{input}, N and S, and reduce D_{outer} and d by 10 times. An example of miniaturization involves decreasing n by 10 times with S fixed. As a result, the acceleration of the magnet increases by ~3.6 times for the same power input, such that the same contact pressure can be achieved by reducing the radial and thickness dimensions of the magnet by a factor of 10 and 3, respectively. The size of other components in the actuator can be adjusted according to the size of the Cu coil.

Device integration and assembly

Low temperature solder joints bonded and electrically connected all of the components, including the SoCs (RF430FRL15xH NFC, Texas Instruments), jumper wires (resistors, 0 Ω, Stackpole Electronics, Inc.), capacitors (10 nF to 2.2 μF, Murata Electronics North America), diodes (SMP1345, Skyworks Solutions, Inc.), power regulator (L78L, STMicroelectronics), IC switches (74LVC1G384, Nexperia BV), jump wires and haptic actuators to corresponding contact pads on the Cu/PI substrate. A thin (~0.2 mm) coating of an ultra-low-modulus silicone material (0.1 mm Silicone C, Silbione, ~3.0 kPa, Ilkem Silicones) on a stretchable fabric substrate (Spandex) served as an adhesive between the cloth and the electronic/haptic platform. Casting and curing a low modulus formulation of PDMS (Silicone B, PDMS, ~60 kPa, Sylgard 184, Dow-Corning) formed a uniform encapsulation layer (2.5 mm thick) over the electrical connections. A layer of skin-coloured PDMS (Silicone A, PDMS, ~0.2 mm, ~60 kPa, Sylgard 184, Dow-Corning) aligned and bonded on top of the device acted as the top encapsulation layer. These soft silicone coatings provided reversible adhesion to the skin, with an adhesion energy of ~90 N m^{-1} for hairless areas and ~80 N m^{-1} for hairy areas (Supplementary Fig. 10). Thin PI (12.5 μm) layers with the same dimensions as the cantilever beams in the haptic actuators were aligned and placed on top of the actuators before covering the top encapsulation layer, enabling the magnets to have freedom for vibratory motions. The total weight is 130 g for a system with 32 actuators in a square array, 40 g for the flower shaped device, 38 g for the oval shaped device, 81 g for the peanut shaped device, 99 g for the triangle shaped device and 120 g for the butterfly shaped device. Supplementary Fig. 11 shows that the water vapour transmission rate (WVTR) of the silicone/fabric sample is ~0.55 g h^{-1} m^{-2}. The WVTR of a similar sample, but with perforating holes (1 mm diameter and 8 mm pitch in a square lattice, with a hole area fraction α of ~1.2%) is ~3.69 g h^{-1} m^{-2}, which is comparable to that of a conventional breathable waterproof bandage (5.72 g h^{-1} m^{-2}) and somewhat smaller than that of a standard cloth bandage g (9.15 g h^{-1} m^{-2}) (Mannings, Hong Kong). We estimate that the complete epidermal VR device, with all of the impermeable active components and interconnects, can accommodate perforations at an overall areal density of $\alpha = 1.2\%$, comparable to that of the test structure. As a result, the WVTR for a system with perforations at this density should also be in the range of 3–4 g h^{-1} m^{-2}. These epidermal VR devices can be worn on the skin (including hairy areas) for extended periods, with various levels of physical activity, without irritation (Supplementary Fig. 12).

Mechanics simulation of the epidermal VR device

The commercial software ABAQUS (v6.10) was used to study the mechanics of the devices. The layouts of the chips and the shapes of the interconnects were optimized to decrease the strain/stress level and to avoid entanglements in the interconnects under different types of external loads (bending, folding and twisting) (Fig. 1m–o). The resonance frequency of the actuator was tuned to 200 Hz (the frequency to which humans are most sensitive) by designing the central angle θ of the PI layer in the actuator (Fig. 1e, Extended Data Fig. 2) to increase the vibration intensity. The arrangement of the actuators was optimized to decrease the mutual interference between them by adjusting their relative angles (Extended Data Fig. 5). The fabric cloth, Silicone A, Silicone B and Silicone C, PDMS, phantom skin and magnet were modelled by hexahedron elements (C3D8R) while the thin copper and PI film were modelled by composite shell elements (S4R). The number of elements in the model was ~3×10^7, and the minimal element size was 1/6 of the width of the narrowest interconnects (50 μm). The mesh convergence of the simulation was guaranteed for all cases. The elastic modulus (E), Poisson's ratio (ν) and density (ρ) are as follows: $E_{cloth} = 391$ kPa, $\nu_{cloth} = 0.4$, $\rho_{cloth} = 0.96 \times 10^3$ kg m^{-3}; $E_{Silicone\_A} = 60$ kPa, $\nu_{Silicone\_A} = 0.5$, $\rho_{Silicone\_A} = 0.96 \times 10^3$ kg m^{-3}; $E_{Silicone\_B} = 60$ kPa, $\nu_{Silicone\_B} = 0.5$, $\rho_{Silicone\_B} = 0.96 \times 10^3$ kg m^{-3}; $E_{Silicone\_C} = 3$ kPa, $\nu_{Silicone\_C} = 0.5$, $\rho_{Silicone\_C} = 0.96 \times 10^3$ kg m^{-3}; $E_{PDMS} = 1$ MPa, $\nu_{PDMS} = 0.5$, $\rho_{PDMS} = 0.96 \times 10^3$ kg m^{-3}; $E_{skin} = 130$ kPa, $\nu_{skin} = 0.5$, $\rho_{skin} = 1.05 \times 10^3$ kg m^{-3}; $E_{magnet} = 113$ GPa, $\nu_{magnet} = 0.34$, $\rho_{magnet} = 8.08 \times 10^3$ kg m^{-3}; $E_{PI} = 2.5$ GPa, $\nu_{PI} = 0.34$, $\rho_{PI} = 0.91 \times 10^3$ kg m^{-3}; and $E_{Cu} = 119$ GPa, $\nu_{Cu} = 0.32$, $\rho_{Cu} = 8.96 \times 10^3$ kg m^{-3}.

Electromagnetic simulation of the epidermal VR device

The finite-element method was used in the electromagnetic simulations to study the magnetic field around the transmission antenna, primary coil, NFC coil, Cu coil (in actuators) and intermediate coil. The simulations were performed using commercial software Ansys HFSS 15 (Ansys Inc.), where the lumped port was used, and the port impedance was set according to the matching capacitor (see below). An adaptive mesh (tetrahedron elements), together with a spherical surface (2,000 mm in radius) as the radiation boundary, were adopted to ensure computational accuracy. The electromagnetic parameters in the material library of Ansys HFSS were used in the simulation.

Characterization of the intermediate coil

The resistance R and inductance L of an intermediate coil (using copper wire 0.7 mm in diameter) with dimensions 20 cm × 20 cm were measured with an impedance analyser. The results agree with simulations (Extended Data Fig. 7e, f). Addition of a matching capacitor with $C = 168$ pF yields a resonant frequency f of ~13.56 MHz (using $f = 1/[2\pi(LC)^{0.5}]$). The Q factor of the intermediate coil, defined by $Q = 2\pi fL/R$, is ~200 at a frequency of 13.56 MHz (Extended Data Fig. 7g).

Electromagnetic simulations for SAR and MPE

The system utilizes RF at 13.56 MHz, a frequency where biological tissues exhibit negligible absorption. The position of the body, therefore, has little effect on the operation, and the system functions well with a variety of obstacles in the environment, even including metal features, owing to the magnetic nature of the wireless link. FEA (using Ansys HFSS 15) was used to determine whether the epidermal VR platform operates within the specific absorbed radiation (SAR) and maximum permissible exposure (MPE) requirements outlined by the Federal Communications Commission (FCC) CFR Part 1.1310. The maximum transmitting power (12 W) of the transmission antenna (852 mm × 620 mm) was adopted in the simulation. The impedance of the circuit part is equivalent to a 170-Ω (the minimal load of the VR system) resistor, measured in the experiments, in series with the primary coil. The distance between the body and the coils/antennas of the epidermal VR system is 3 mm, which is the distance when the epidermal VR system is mounted on the body. The density of the body is taken as 1,000 kg m^{-3}.

NFC protocols, software control and system operation

Programming each SoC (RF430FRL15xH, Texas Instruments) with the Code Composer Studio (CCS) enabled the generation of square wave signals by alternating the output of the GP I/O port between its HIGH (high voltage output) and LOW (zero voltage output) settings. The frequency of the resulting square waves can be adjusted from 100 Hz to 300 Hz in a user-definable way through the software interface through changes in the SoC program. Each SoC controls eight separate GP I/O ports, and each of these can be controlled independently. The program incorporated an interrupt mechanism as the control on and off for the square wave. Writing a specific hexadecimal value into a particular register on the SoC activated the GP I/O port to generate the square wave signal. Writing any other hexadecimal value into this same register deactivated the port. As such, the interrupt acts as the control mechanism for the square wave. Other hexadecimal commands initiated data transfer via NFC Data Exchange Format (NDEF) messages. The necessary NDEF messages can be written into the SoC with a transmission antenna that operates at 13.56 MHz (FEIG, ID ISC. LRM2500-A), with an output power between 1 W and 12 W and the ability to interface to a computer/laptop via a USB port. The reader, connected to a computer/laptop, in this manner served as an interface to control the writing process using a custom graphical user interface. The software interface displays the connection status of the reader with the computer. Each virtual button (pixel) on the touch panel is associated with a 4-byte command. Control of each actuator can, in this way, be achieved without any interference or cross-talk. To avoid time delay, sequential operation was programmed into the SoC, leading to millisecond response times. SoCs can be differentiated from one another by their unique ID numbers (Supplementary Fig. 5). In this way, control of multiple SoCs can be accomplished without uncertainties in directionality in RF transmission.

Visualization of system level operation using an array of 32 lasers

The measurements relied on a custom-built system with an array of 32 lasers. Each laser beam (635 nm, 2 mW continuous wave) reflects from a small reflective disk (0.15-mm-thick pieces of glass of diameter 8 mm, coated with 10 nm Cr and 100 nm Au) mounted on a corresponding haptic actuator in the epidermal VR system. The full collection of reflected beams projects onto a monitoring screen to allow direct visualization of the spatial patterns of activation, as well as the amplitudes and orientations of the vibratory motions of the cantilevers associated with the actuators. The long light path geometrically converts the small amplitude vibrations of the actuators into large motions of the reflected spots, for easy visualization. An optical system of mounting stages allowed adjustment of each laser beam, including its incident angle, tilt angle and distance from the actuators. Each of the 32 laser modules was held by an angle post clamp, enabling an independently adjustable laser pattern to be configured. Once a laser module was adjusted to the proper position, the angle post clamps were fixed.

In this customized system, six horizontal posts supported the 32 laser modules: two of the posts supported four modules, the rest each supported six modules. The six posts were arranged one above the other, and this array was mounted on two vertical posts by angle post clamps. The two vertical posts were fixed on a solid aluminium optical breadboard (Extended Data Fig. 9a). Careful adjustments yielded a pattern of laser beams that were aligned with all the haptic actuators across the device.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data are contained within the manuscript. Raw data are available from the corresponding authors upon reasonable request.

Code availability

Custom code used in this study is available from the corresponding authors upon reasonable request.

23. Lee, Y. *RFID Coil Design* (Microchip Technology Inc., 1998).

Acknowledgements We gratefully acknolwedge C. J. Su, T. Banks, J. H. Kim, Y. G. Xue and J. K. Chang for their efforts in constructing and testing the optimized systems. This work was supported by the Center for Bio-Integrated Electronics at Northwestern University. Z.X. and X.F. acknowledge support from the National Basic Research Program of China (grant number 2015CB351900) and the National Natural Science Foundation of China (grant numbers 11320101001 and 11402134). X.Y., Z.X., D.L. and Y.L. acknowledge support from the City University of Hong Kong (grant number 9610423). R.S. acknowledges support from the Engineering and Physical Sciences Research Council (grant number EP/L016028/1) and the China Scholarship Council. Y.H. acknowledges support from the NSF (grant number 1635443).

Author contributions X.Y., Z.X., Y.Y. and J.L. contributed equally to this work. X.Y, Z.X., H.Y. and J.A.R. conceived the ideas, designed the experiments and wrote the manuscript. X.Y., Z.X., J.L., J.R., A.V., H.L., X.N., A.A., D.L., B.J., R.S., Y.L., J.C., Q.H., C.L., S.K., P.G., C.Z., Y.Z., Y.X., Q.G., A.C., P.T., W.L., J.J., Y.Y., J.C., C.T., B.K., K.L., X.F. and J.C. performed experiments and analysed the experimental data. Z.X. led the structural designs and performed electromagnetic and mechanical modelling, with assistance from B.J., Q.H., C.Z. and Y.X.

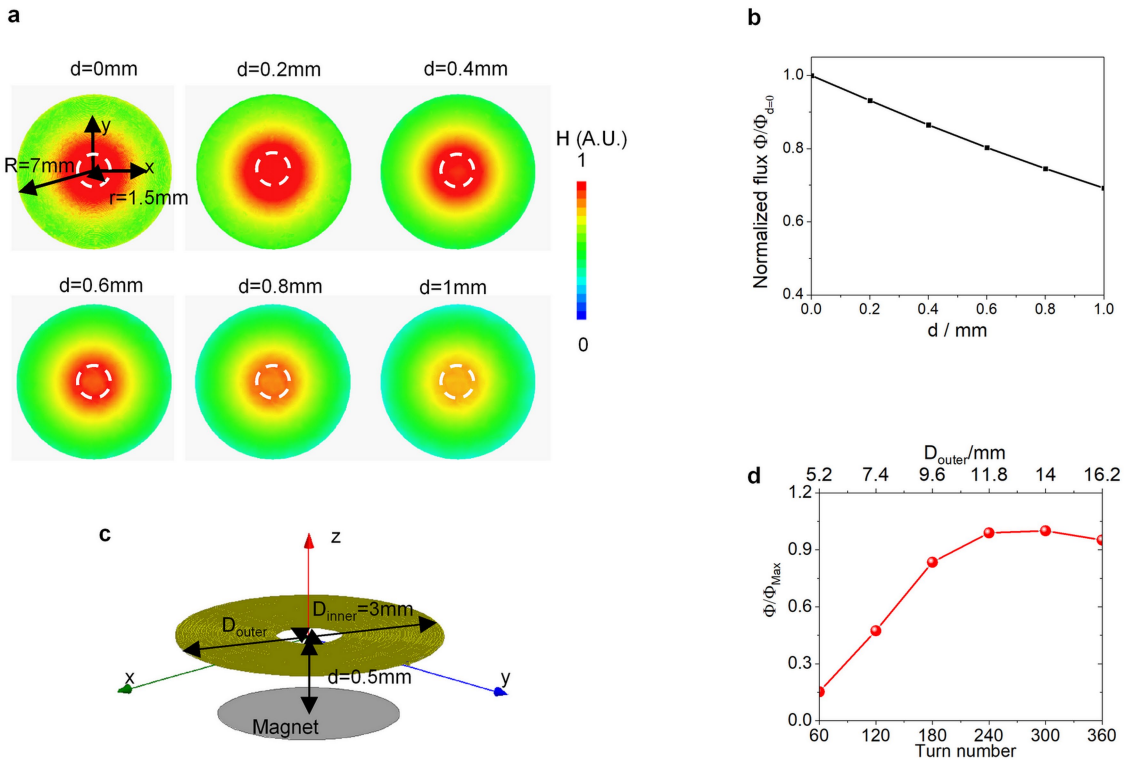
Competing interests The authors declare no competing interests.

Additional information

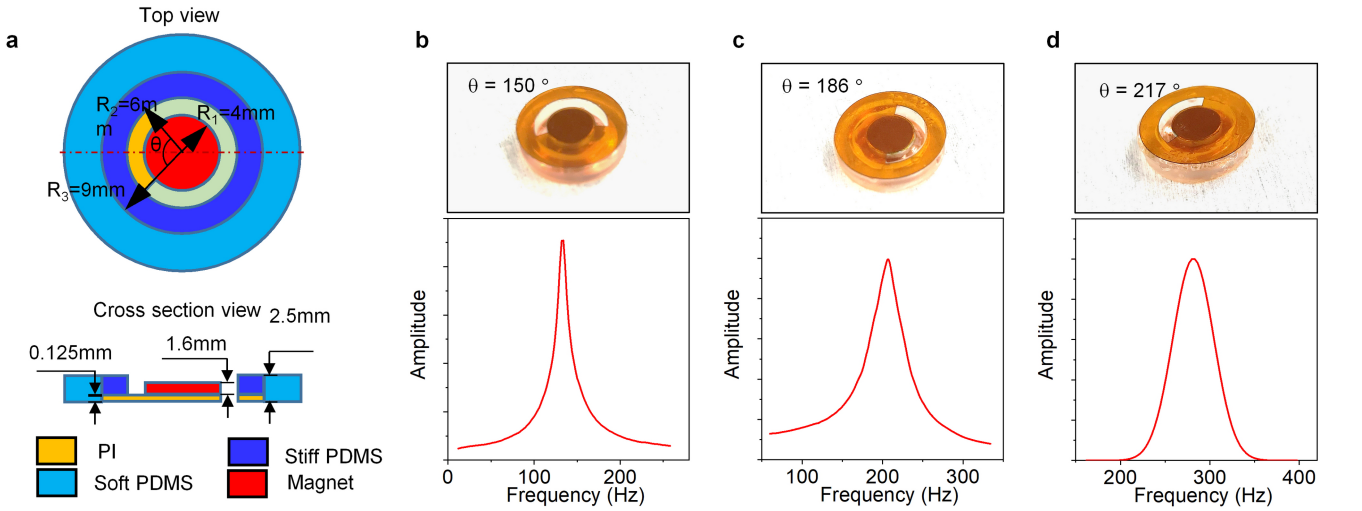
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1687-0.

Correspondence and requests for materials should be addressed to Y.H. or J.A.R.

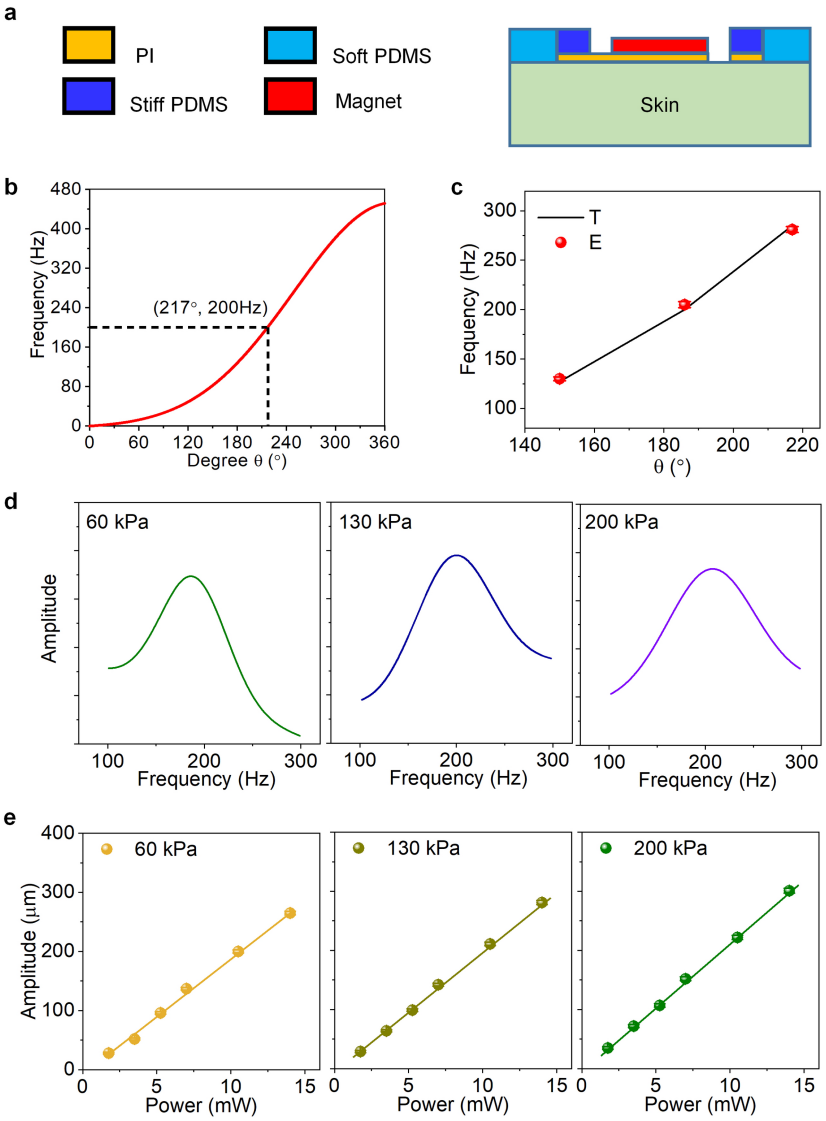
Peer review information *Nature* thanks Sriram Subramanian, Xiaoming Tao and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



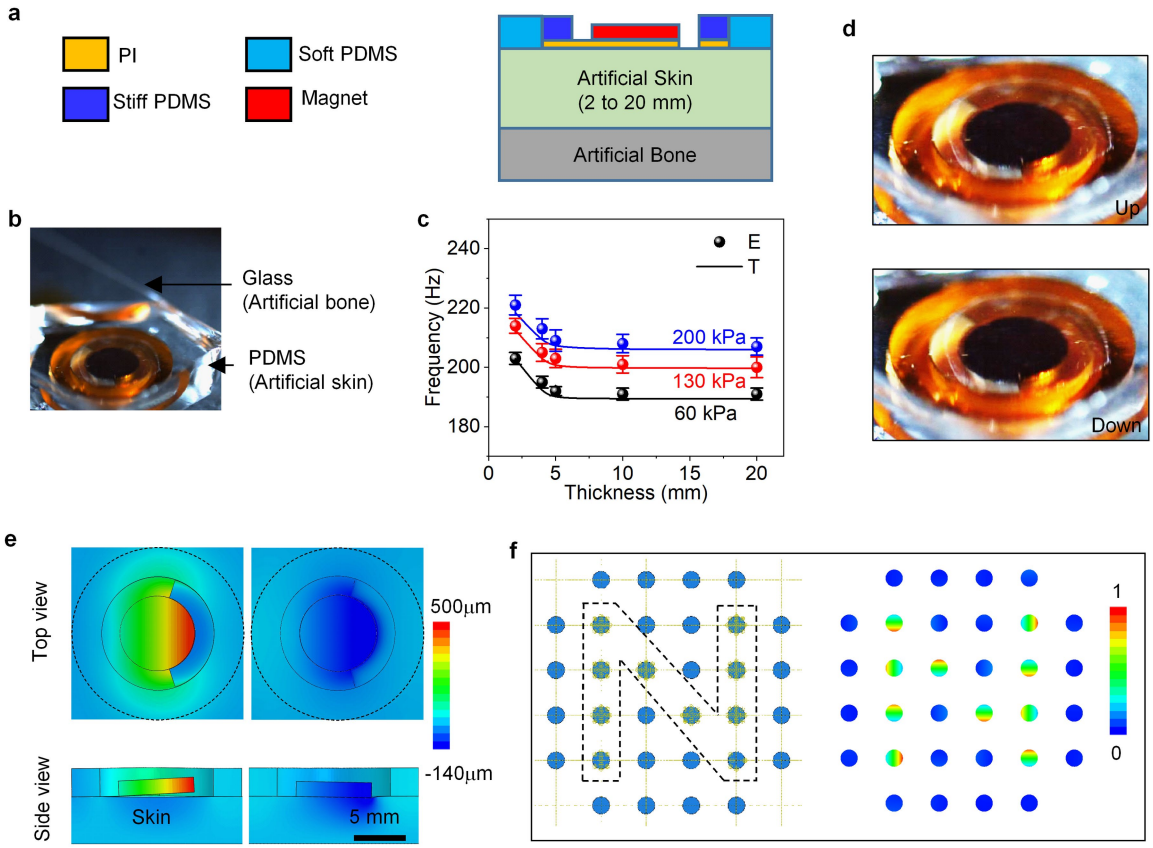
Extended Data Fig. 1 | Study of the magnetic field strength in an actuator. a, Distribution of the normalized magnetic field strength (H) around the Cu coil (300 turns) used in the actuator, where d is the distance between the Cu coil and the magnet. The dashed white circles correspond to the holes in the Cu coil. **b**, Normalized magnetic flux ($\Phi/\Phi_{d=0}$) through the haptic actuator versus the distance to the Cu coil (300 turns) of the actuator. **c**, Schematic illustration of the distance between a haptic actuator and a Cu coil in the actuator. **d**, The normalized magnetic flux (Φ/Φ_{Max}) through the magnet versus turn number and outer diameter (D_{outer}) of the Cu coil.



Extended Data Fig. 2 | Resonant frequency tuning of the actuator. a, Top view and cross-sectional view of the actuator design. The parameters presented here are optimized for tuning the resonance frequency to the skin-sensitive range. **b**–**d**, Optical images and normalized amplitude–frequency curves of three actuators with different central angles (θ) of 150° (**b**), 186° (**c**) and 217° (**d**), working without any contact. The actuators shown in **b** and **c** are 18 mm in diameter and 2.5 mm thick.

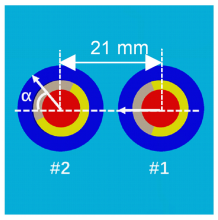


Extended Data Fig. 3 | Optimization of the actuator. a, Cross-sectional schematic illustration of an actuator in contact with skin. **b**, Theoretical results of the resonance frequency of actuators shown in **a** versus the central angle θ of the PI handling layer. The dashed lines indicate the resonance frequency of 200 Hz at $\theta = 217°$. **c**, Comparison of experimental (E; data symbols) and simulation (theory, T; lines) results. **d**, Experimental results of the normalized amplitude–frequency curves of the actuator ($\theta = 217°$) in contact with skin, for different values of skin elastic modulus: 60 kPa, 130 kPa and 200 kPa. **e**, Travel amplitude of the magnet as a function of the input power (data points) for an actuator in contact with artificial skin samples with elastic moduli of 60 kPa, 130 kPa and 200 kPa. In **c**, **e**, error bars correspond to the calculated standard deviation.

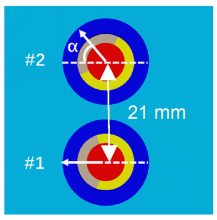


Extended Data Fig. 4 | Study of the mechanical behaviour of the actuator when in contact with skin. a, **b**, Cross-sectional schematic illustration (**a**) and optical image (**b**) of an actuator in contact with artificial skin. Here PDMS, serving as artificial skin, had three different values of elastic modulus, and glass was used for artificial bone. **c**, Comparison of the experimental (E) and simulation (theory, T) results of the resonance frequency of the actuator in contact with different moduli and thicknesses of artificial skin. The error bars correspond to the calculated standard deviation. **d**, Optical images captured using a high-speed camera and a working actuator travelling up and down, when in contact with 130-kPa artificial skin. **e**, FEA results for the amplitude of an actuator: left, when separated from skin; right, when in contact with skin. **f**, Schematic illustration and FEA results (colour-coded amplitude) for mechanical coupling between an array of haptic actuators, with activation in an 'N' pattern.

a

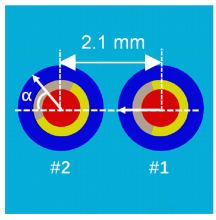


α (°)	Amplitude ratio
0	5.9%
45	2.8%
90	5.9%
135	8.0%
180	7.5%
225	8.0%
270	5.9%
315	2.8%

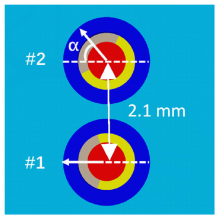


α (°)	Amplitude ratio
0	9.9%
45	5.8%
90	3.3%
135	9.4%
180	15.1%
225	11.3%
270	5.7%
315	7.4%

b



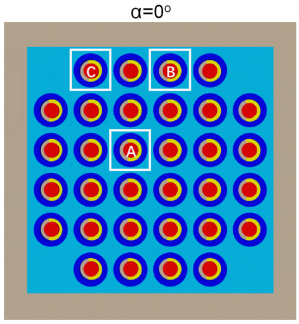
α (°)	Amplitude ratio
0	3.7 %
45	1.2 %
90	3.9 %
135	5.4 %
180	4.6 %
225	5.4 %
270	3.9 %
315	1.2 %



α (°)	Amplitude ratio
0	6.3 %
45	3.4 %
90	1.1 %
135	3.6 %
180	6.6 %
225	4.2 %
270	1.3 %
315	5.2 %

c

α=0°

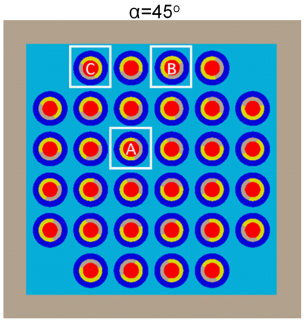


	0.03	0.03	0.04	0.03	
0.02	0.02	0.15	0.02	0.03	0.03
0.01	0.01	1.00	0.01	0.02	0.02
0.02	0.02	0.14	0.02	0.02	0.02
0.02	0.02	0.02	0.03	0.02	0.02
	0.02	0.02	0.02	0.02	

	0.01	0.01	1.00	0.02	
0.03	0.03	0.02	0.14	0.02	0.03
0.03	0.03	0.04	0.03	0.03	0.03
0.02	0.02	0.02	0.02	0.02	0.01
0.01	0.01	0.01	0.00	0.00	0.01
	0.02	0.02	0.02	0.02	

	1.00	0.01	0.01	0.02	
0.01	0.13	0.02	0.03	0.03	0.03
0.02	0.02	0.03	0.03	0.03	0.03
0.01	0.02	0.02	0.01	0.01	0.01
0.01	0.01	0.01	0.01	0.01	0.01
	0.02	0.02	0.02	0.01	

α=45°

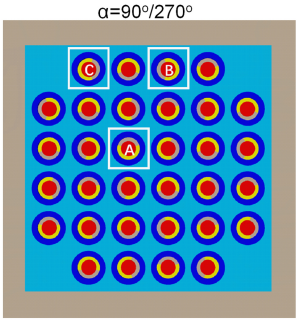


	0.03	0.01	0.04	0.03	
0.02	0.02	0.06	0.03	0.02	0.02
0.01	0.02	1.00	0.03	0.01	0.03
0.02	0.02	0.06	0.01	0.03	0.03
0.03	0.02	0.01	0.03	0.03	0.02
	0.01	0.02	0.03	0.01	

	0.03	0.05	1.00	0.01	
0.04	0.03	0.05	0.06	0.03	0.03
0.03	0.02	0.04	0.04	0.02	0.05
0.02	0.02	0.03	0.01	0.03	0.02
0.01	0.02	0.02	0.01	0.01	0.02
	0.02	0.02	0.02	0.02	

	1.00	0.06	0.03	0.03	
0.01	0.01	0.02	0.03	0.02	0.02
0.02	0.01	0.03	0.03	0.01	0.03
0.02	0.01	0.03	0.01	0.01	0.01
0.01	0.02	0.02	0.01	0.01	0.02
	0.01	0.02	0.02	0.02	

α=90°/270°



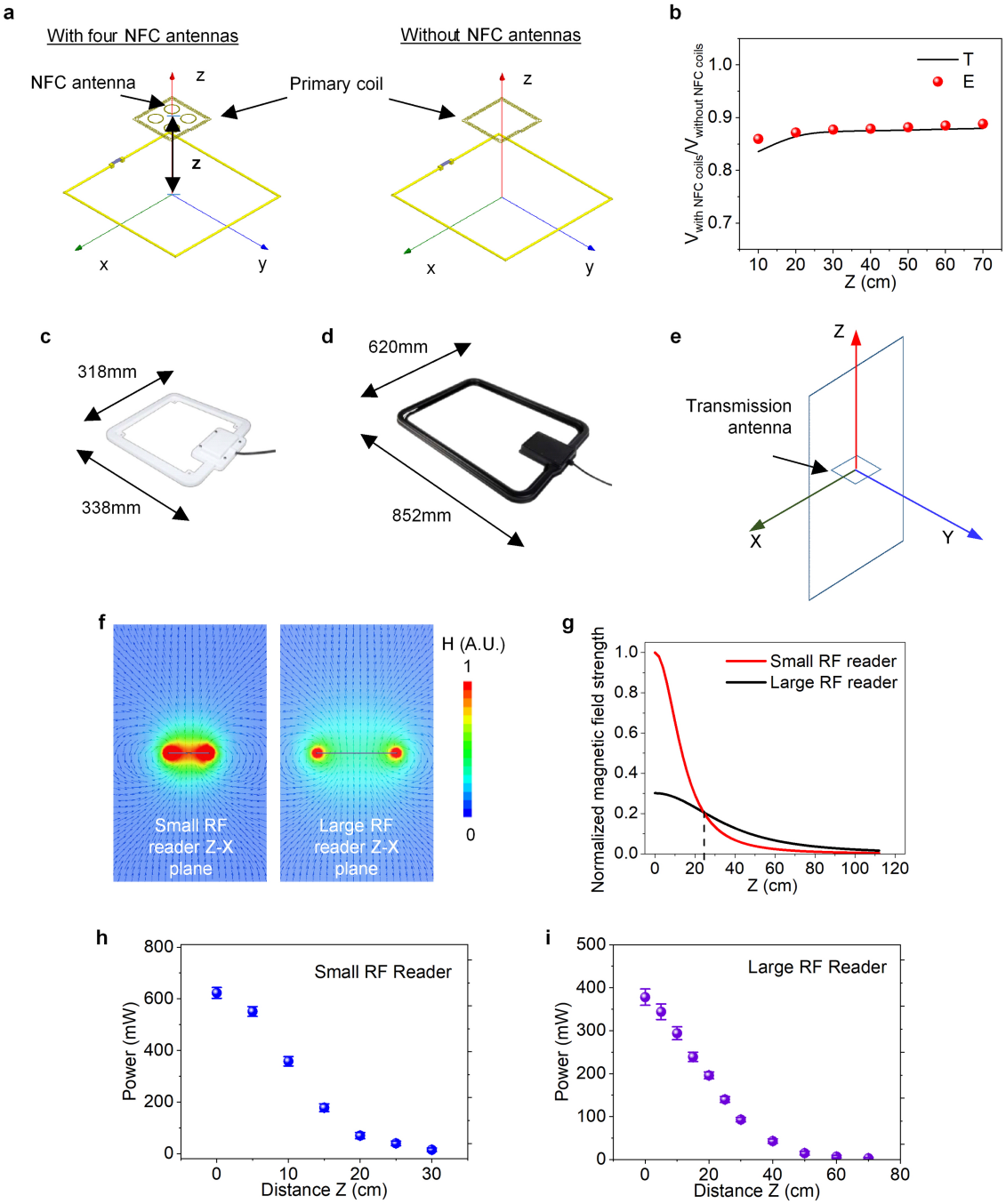
	0.01	0.02	0.01	0.03	
0.01	0.04	0.05	0.03	0.02	0.03
0.04	0.03	1.00	0.03	0.04	0.01
0.02	0.03	0.04	0.04	0.02	0.01
0.04	0.01	0.03	0.01	0.04	0.01
	0.03	0.01	0.03	0.02	

	0.01	0.06	1.00	0.04	
0.02	0.02	0.03	0.06	0.01	0.01
0.02	0.03	0.02	0.04	0.01	0.03
0.02	0.03	0.03	0.01	0.03	0.02
0.03	0.01	0.02	0.01	0.01	0.01
	0.02	0.01	0.00	0.00	

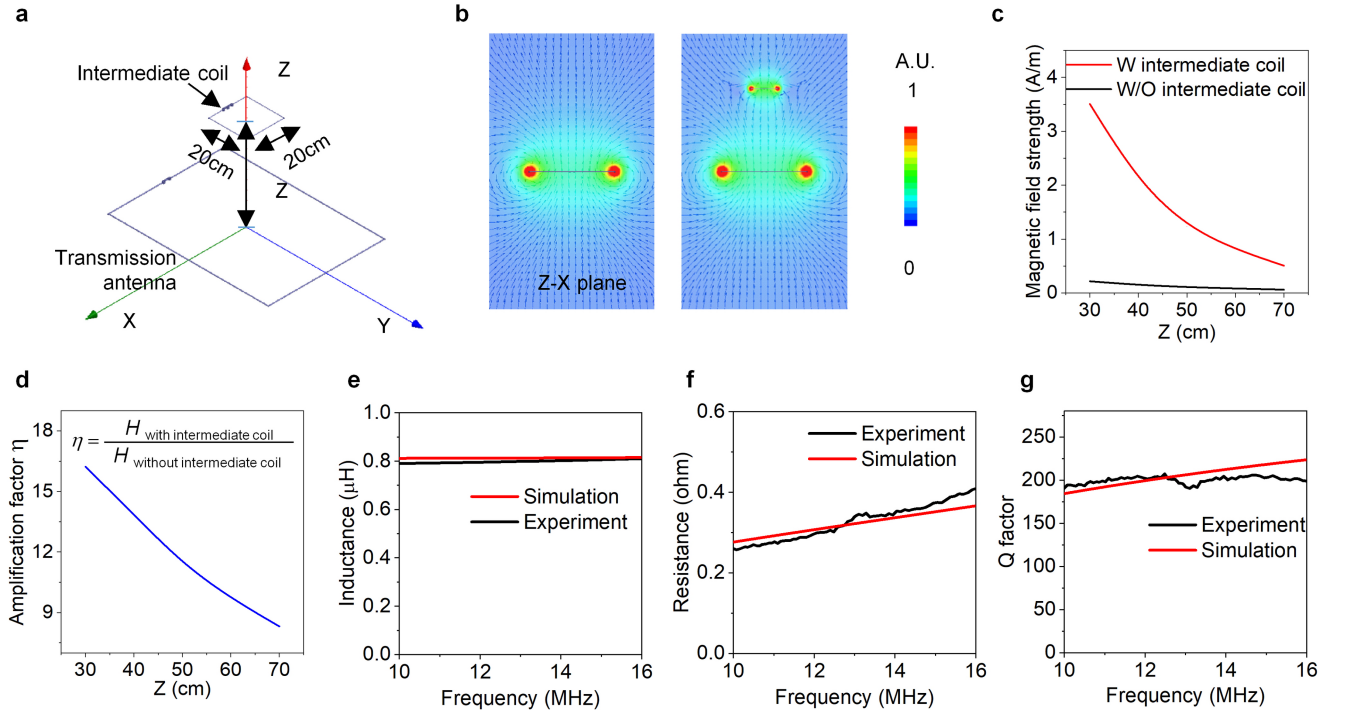
	1.00	0.04	0.01	0.01	
0.01	0.01	0.01	0.01	0.03	0.02
0.01	0.03	0.01	0.03	0.02	0.03
0.02	0.02	0.03	0.01	0.03	0.03
0.01	0.01	0.01	0.01	0.02	0.01
	0.01	0.00	0.01	0.01	

Extended Data Fig. 5 | See next page for caption.

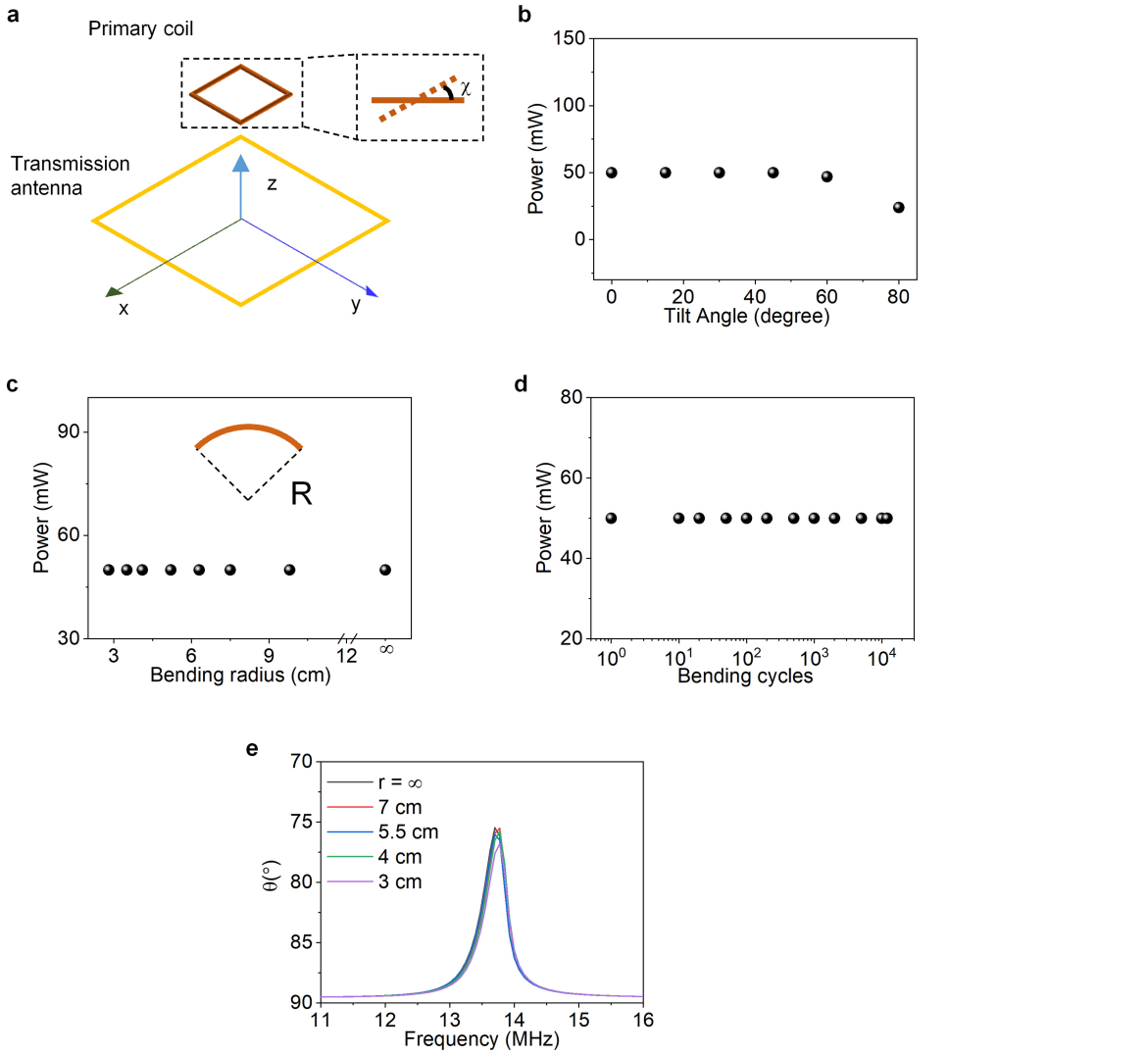
Extended Data Fig. 5 | Mutual interference study of an actuator array. a, Mutual interference of two actuators at different relative angles α at 200 Hz. Two representative cases were studied, with one actuator (no. 1) positioned along (left) or perpendicular to (right) the bisector of the other actuator (no. 2). Here, only actuator 1 was actuated. The amplitude ratio, that is, the amplitude of actuator 2 over the amplitude (induced by mutual interference) of actuator 1, shown in the table demonstrates that $\alpha = 45°$, 90° and 270° result in relatively small mutual interference for both representative cases simultaneously. **b**, Mutual interference of two small actuators at different relative angles α at their resonant frequency of 200 Hz. The size of the actuators and the distance between them were scaled to 1/10 of the original design as shown in **a**, that is, the distance between two actuators was 2.1 mm rather than 21 mm. Here the thickness of the PI disk was set at 1.8 μm to enable the resonant frequency for the small actuator to be 200 Hz. Two representative cases were studied, with one actuator (no. 1) positioned along (left) or perpendicular to (right) the bisector of the other actuator (no. 2). Here only actuator 1 was actuated. The amplitude ratios (actuator 2 to actuator 1) due to mutual interference are shown in the table. **c**, Optimization of the actuators' arrangement. The mutual interference among actuators was studied for $\alpha = 0°$, $\alpha = 45°$ and a combination of 90°/270° referring to the simulation results in **a**, and for representative cases when actuators A (around the centre), B (near the boundary) and C (near the corner) are actuated separately. The results show that $\alpha = 90°/270°$ yields the smallest mutual interference (see tables under). The number gives the amplitude ratio due to the mutual interference among the actuators—that is, the amplitude of all actuators over the amplitude of the activated actuator—where 1 represents the activated actuator.



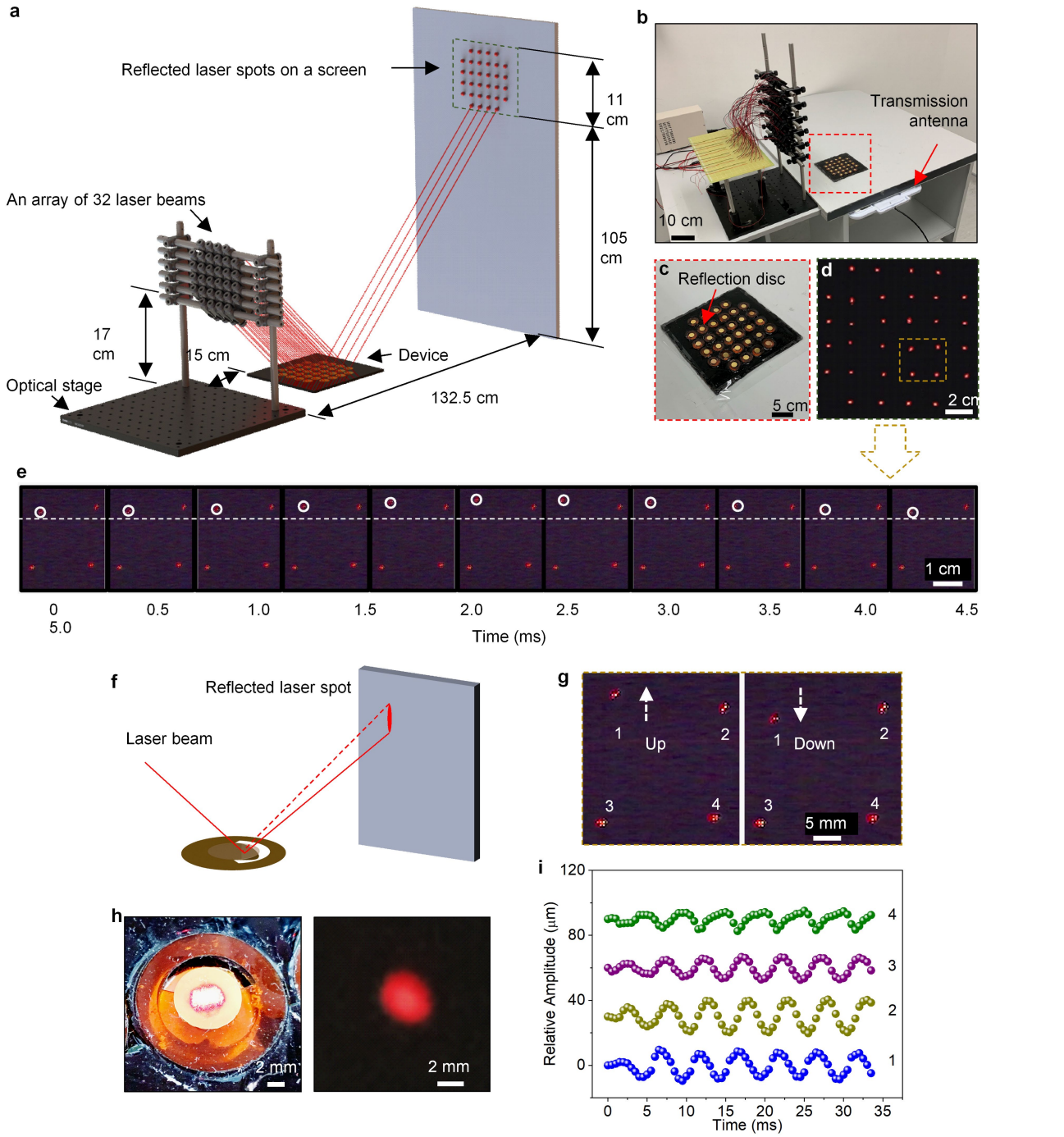
Extended Data Fig. 6 | Study of the key electrical components of the epidermal VR system. a, Interference between the primary coil and NFC antennas. Shown are schematic illustrations of a primary coil with (left) or without (right) four NFC antennas along the *Z* direction of the transmission antenna. **b**, Comparison of experimental (E) and theoretical (T) results of voltage induced by a single primary coil versus a primary coil with four NFC antennas. **c**, **d**, Transmission antennas used for operating the VR devices: **c**, small size, 318 mm × 338 mm; and **d**, large size, 620 mm × 852 mm. **e**, An transmission antenna placed in the *X*–*Y* plane. **f**, The magnetic field strength (*H*) in the *Z*–*X* plane (the middle plane of the coil) for the small (left) and large (right) transmission antennas (RF readers). **g**, Theoretical results show that the small and large transmission antennas are suitable for short (<24 cm) and long (>24 cm) working distances, respectively. **h**, **i**, Output power of the serpentine primary coil of the epidermal VR device as a function of distance to the small transmission antenna (**h**), and the large transmission antenna (**i**). The error bars correspond to the calculated standard deviation.



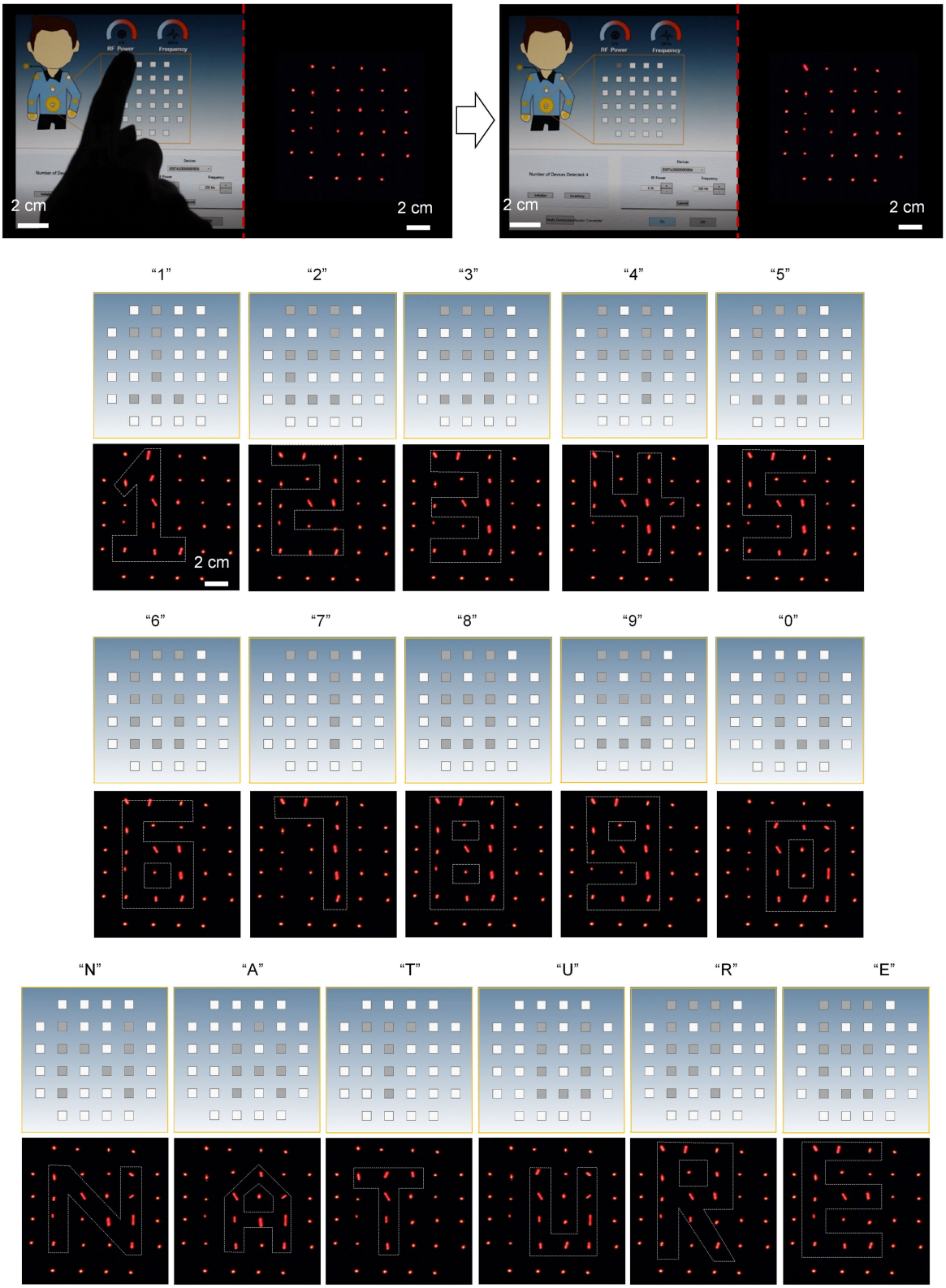
Extended Data Fig. 7 | Electrical properties of the intermediate coil. **a**, Configuration of an intermediate coil (20 cm × 20 cm, wound with Cu wire with a diameter of 0.2 mm) oriented parallel to the *X*–*Y* plane, at a distance *Z* from the transmission antenna. **b**, Computational results for the magnetic field distribution induced by an RF transmission antenna tuned to resonance with a receiver antenna with and without an intermediate coil. **c**, Comparison of the magnetic field strength along the *Z* direction of the transmission antenna with (W) and without (W/O) the intermediate coil. **d**, Amplification factor η along the *Z* direction of the transmission antenna with and without the intermediate coil. **e**–**g**, Simulation and experimental results for the inductance (**e**), resistance (**f**) and Q factor (**g**) as a function of frequency.



Extended Data Fig. 8 | Mechanical characterization of the epidermal VR system. a–d, Output power from the primary coil of an epidermal VR device as a function of tilt angle χ (**a**, geometry; **b**, data), bending radius, R (**c**) and bending cycles to an R of 2.8 cm (**d**). The distance between the device and the antenna was fixed at 20 cm for all measurements. **e**, Measured phase responses of the antennas used for wireless control over the SoCs as a function of radius of curvature. The resonance frequency is 13.56 MHz before bending. Bending induces only very slight shifts in these curves.



Extended Data Fig. 9 | Visualization of system level operation using an array of 32 lasers. a, Schematic illustration of a custom-built laser array system for real-time visualization of the operation of a complete epidermal VR system. **b**–**d**, The array of lasers (**b**), the corresponding array of beams reflecting from the haptic actuators, each mounted with a reflective disk (diameter of 8 mm), across an entire system (**c**), and their arrival at a monitoring screen (**d**). **e**, Representative frames extracted from video recorded using a high-speed camera to capture oscillatory motions of each of the laser spots. These motions directly determine the motions of the cantilever-based actuators. **f**, Schematic illustration of a laser spot produced by projection of a reflected beam onto a screen during the operation of the actuator. **g**, Representative frames extracted from video recorded using a high-speed camera showing the oscillatory motion of laser spot 1. **h**, Pictures of a laser spot on a reflector that mounts on a haptic actuator (left) and on the monitoring screen (right). The diameter of the laser spot is ~3 mm. **i**, Calculated displacements of four actuators determined from the measurement setup geometry and the amplitude of motion of the laser spots in **g**, **e** and Supplementary Video 4. The traces are offset in the *y* direction to facilitate visual inspection. The calculated displacements are somewhat smaller than those measured directly from individual actuators using high-speed cameras owing to slight misalignments of the lasers and to shifts in the resonance frequencies due to absence of the PDMS encapsulation layer for the devices measured using the laser technique. The results allow direct visualization and measurement of the vibration amplitudes, direction and frequency of the cantilever beams associated with each actuator across the full array.



Extended Data Fig. 10 | Pictures of the operation of an epidermal VR system, visualized with a reflected array of 32 laser beams. Activation of a given haptic actuator causes the corresponding reflected spot to transform from a circular to an elliptical shape, owing to the vibratory motions (top row). The results in the lower three rows show representative spatial patterns of actuation, including numbers 0 to 9 and letters 'N', 'A', 'T', 'U', 'R' and 'E'. We note that the detailed shapes of the laser spots on the screen depend critically on the positioning of each of the beams across the corresponding reflectors mounted on the cantilevered actuator structures.

nature research

Corresponding author(s): John A. Rogers

Last updated by author(s): Jun 12, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Keysight E4991B Impedance Analyzer, Labview 2016, RSR MS8229 digital multimeter, Powerlab by AD instrument.
Data analysis	MATLAB R2017b, Powerlab by AD instrument, Excel 2016, Origin 2018.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data is contained within the manuscript. Reasonable requests can be made to the corresponding author for access to the raw data.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were preliminary estimated by finite element analysis using software "Abaqus", then determined by preliminary experiments.
Data exclusions	No data were excluded.
Replication	All experiments were performed for at least 5 times. All attempts at replication were successful.
Randomization	All unit devices (actuators) were fabricated and tested first to make sure they are functional. Then integration all good devices and electronic components together to fabricate the system level devices.
Blinding	All data was collected in an unbiased manner and experiments was blinded to treatment for final analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Real-space charge-density imaging with sub-ångström resolution by four-dimensional electron microscopy

https://doi.org/10.1038/s41586-019-1649-6

Received: 7 November 2018

Accepted: 6 August 2019

Published online: 14 October 2019

Wenpei Gao[1,8], Christopher Addiego[2,8], Hui Wang[2,7,8], Xingxu Yan[1], Yusheng Hou[2], Dianxiang Ji[3], Colin Heikes[4], Yi Zhang[1], Linze Li[1], Huaixun Huyan[1], Thomas Blum[2], Toshihiro Aoki[5], Yuefeng Nie[3], Darrell G. Schlom[4,6], Ruqian Wu[2]* & Xiaoqing Pan[1,2,5]*

The distribution of charge density in materials dictates their chemical bonding, electronic transport, and optical and mechanical properties. Indirectly measuring the charge density of bulk materials is possible through X-ray or electron diffraction techniques by fitting their structure factors[1–3], but only if the sample is perfectly homogeneous within the area illuminated by the beam. Meanwhile, scanning tunnelling microscopy and atomic force microscopy enable us to see chemical bonds, but only on the surface[4–6]. It remains a challenge to resolve charge density in nanostructures and functional materials with imperfect crystalline structures—such as those with defects, interfaces or boundaries at which new physics emerges. Here we describe the development of a real-space imaging technique that can directly map the local charge density of crystalline materials with sub-ångström resolution, using scanning transmission electron microscopy alongside an angle-resolved pixellated fast-electron detector. Using this technique, we image the interfacial charge distribution and ferroelectric polarization in a $SrTiO_3$/$BiFeO_3$ heterojunction in four dimensions, and discover charge accumulation at the interface that is induced by the penetration of the polarization field of $BiFeO_3$. We validate this finding through side-by-side comparison with density functional theory calculations. Our charge-density imaging method advances electron microscopy from detecting atoms to imaging electron distributions, providing a new way of studying local bonding in crystalline solids.

The ability to directly visualize the distribution of electrons in solids and molecules could greatly advance science, as nearly all physical properties of materials are determined by the rearrangement of electron charges between nuclei when atoms aggregate together. However, observing electrons in materials at high spatial resolution is not routine. Unlike other diffraction methods[1–3], aberration-corrected scanning transmission electron microscopy (AC-STEM) offers the possibility of achieving atomic-resolution imaging, by using an electron beam focused to a sub-ångström width. While penetrating through a specimen, the electron beam interacts with the local electric field in its pathway, resulting in a change in its momentum. Recently, differential phase contrast (DPC) imaging in STEM was developed to estimate the momentum of the electron probe by counting the integrated scattering electrons in segments of an annular STEM detector[7,8]. If the pattern of the entire electron beam can be captured using a fast camera in STEM, the electric field at each scanned position can be calculated from the momentum change with high precision. Using this method, electric fields and charge densities have been mapped in simulations[9], but experimental results have been stymied by strong noise and insufficient resolution[10]. Here, using the state-of-the-art AC-STEM and a high-speed pixellated electron detector, we have successfully imaged the local electric field and charge density in $SrTiO_3$, $BiFeO_3$ and a $SrTiO_3$/$BiFeO_3$ heterojunction with ultrahigh spatial resolution in real space. The high fidelity of this approach—supported by comparisons with density functional theory (DFT) calculations—shows its potential for studying the complex interplay between charge, field and atomic structure in heterogeneous materials.

We first carried out real-space charge-density imaging (RSCDI) of $SrTiO_3$ by performing four-dimensional (4D) scanning electron diffraction using an AC-STEM equipped with a high-speed pixellated electron detector (Gatan OneView; Fig. 1a, b and Methods). $SrTiO_3$ is a simple cubic perovskite with a projected square symmetry (Fig. 1c). In Fig. 1d, the high-angle annular dark field (HAADF)-STEM image of $SrTiO_3$ clearly shows the Sr and Ti atomic columns, whereas O atoms cannot be seen

[1]Department of Materials Science and Engineering, University of California at Irvine, Irvine, CA, USA. [2]Department of Physics and Astronomy, University of California at Irvine, Irvine, CA, USA. [3]National Laboratory of Solid State Microstructures, College of Engineering and Applied Sciences, Collaborative Innovation Center of Advanced Microstructures, Nanjing University, Nanjing, China. [4]Department of Materials Science and Engineering, Cornell University, Ithaca, NY, USA. [5]Irvine Materials Research Institute, University of California at Irvine, Irvine, CA, USA. [6]Kavli Institute at Cornell for Nanoscale Science, Cornell University, Ithaca, NY, USA. [7]Present address: Hunan Key Laboratory of Super Microstructure and Ultrafast Process, School of Physics and Electronics, Central South University, Changsha, China. [8]These authors contributed equally: Wenpei Gao, Christopher Addiego, Hui Wang. *e-mail: wur@uci.edu; xiaoqing.pan@uci.edu

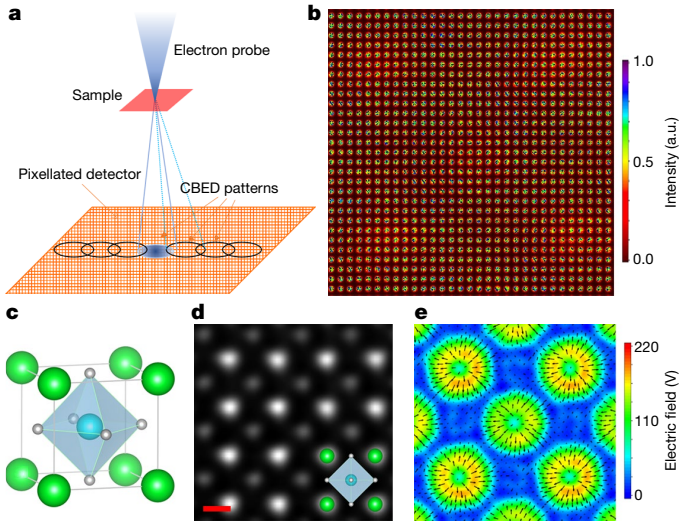


Fig. 1 | Experimental setup and the electric field in $SrTiO_3$. a, The experimental setup for 4D scanning electron diffraction using an AC-STEM. As an electron probe scans across the two-dimensional (2D) surface of a sample, a 2D diffraction pattern at each point is acquired using a pixellated detector, and a convergent beam electron diffraction (CBED) pattern is formed, enabling further analysis. **b**, The as-acquired scanning diffraction dataset from a unit cell in $SrTiO_3$. a.u., arbitrary units. **c**, Atomic structure obtained from DFT calculations. **d**, HAADF-STEM image of $SrTiO_3$. Scale bar, 2 Å. **e**, Corresponding electric-field map derived using the shift in the centre of mass from the scanning diffraction dataset. The vector arrows represent the direction and magnitude of the local electric field (in volts) in a unit cell of $SrTiO_3$.

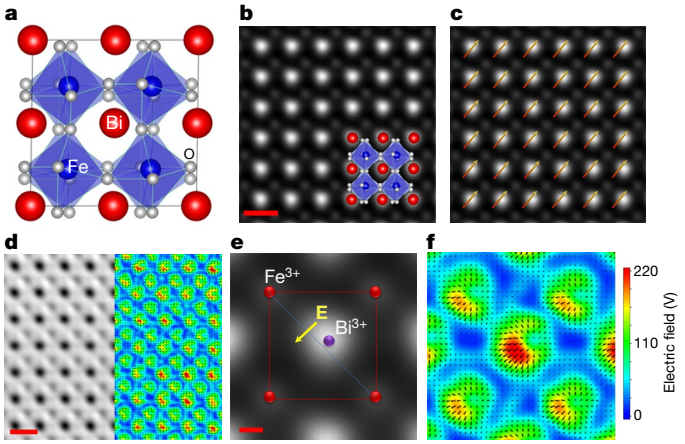


Fig. 2 | Atomic structure and electric-field dipole of $BiFeO_3$. a, Atomic structure of $BiFeO_3$, obtained from DFT calculations. Bi atoms are in red, Fe in blue and O in grey. **b**, Atomic-resolution HAADF-STEM image of $BiFeO_3$; scale bar, 4 Å. The Bi atomic columns are seen shifted towards the top right of each unit cell (defined by the four Fe columns in the corners). **c**, The dipole moment is polarized towards the top right, as shown by the vectors. **d**, Conventional STEM bright-field image (left) and electric-field map (right) of $BiFeO_3$, generated from scanning diffraction data. Scale bar, 4 Å. **e**, Magnified reconstructed dark-field STEM image of a $BiFeO_3$ unit cell from the as-acquired scanning diffraction dataset. Bi is marked in purple, Fe in red. The direction of the electric field (**E**) is shown schematically. Scale bar, 1 Å. **f**, Corresponding electric-field map, where the colour represents the magnitude of the local electric field. The electric-field vector around Bi is stronger in the direction of the bottom left corner. Around Fe columns, the electric field is also deformed diagonally towards the bottom left.

owing to their weak scattering of electrons. In the 4D dataset (Fig. 1b), regions surrounding the Sr and TiO columns are brighter, as electrons are scattered more strongly by heavy nuclei. Capturing the entire diffraction pattern allows us to calculate the shift in the centre of the diffraction intensity (also referred to as the centre of mass, COM) at each point of the electron probe, which covers an area of 0.6 Å in diameter. The lateral shift, $\Delta\mathbf{COM}$, relates to the change in the momentum of the electron beam, $\Delta\mathbf{p}_{xy}$, and is negatively proportional to the local electric field, $\mathbf{E}_{xy}$, when the sample is thin (Extended Data Fig. 1)[9,10]. Explicitly, the electric field can be calculated using:

$$\mathbf{E}_{xy} = \frac{-\Delta\mathbf{p}_{xy}}{e}\frac{v_z}{\Delta z} = -\frac{\Delta\mathbf{COM}p_z}{e}\frac{v_z}{\Delta z}$$

where Δz is the sample thickness, v_z is the speed of the electrons along the beam direction, e is the electron charge and p_z is the linear momentum of electrons along the beam direction.

Figure 1e shows the electric-field map derived from the 4D dataset for a $SrTiO_3$ sample with a thickness of 5.6 ± 1 nm (the sample thickness was measured by least-squares fitting of the experimental position-averaged convergent beam electron diffraction (PACBED)[11] data with the simulated datasets; see Extended Data Figs. 2, 3 for details). The vectors show the direction and magnitude of the local electric field. Note that the electric field is radially distributed around each Sr, Ti and O atom, suggesting that it is highly symmetrical around the ions in $SrTiO_3$. Furthermore, it is obvious that the electric field around the O columns is much weaker than that around the Sr and TiO columns.

In contrast to $SrTiO_3$, $BiFeO_3$ adopts a rhombohedral phase at room temperature[12] (Fig. 2a). DFT calculations show that electric polarization in $BiFeO_3$ originates from the deformation of the Bi 6*s* lone-pair electrons, which drives the rotation of FeO octahedra along the (111) axis and displaces the Fe atoms away from the centre of the surrounding oxygen's octahedral cage[13]. Figure 2b shows an HAADF image along the $BiFeO_3$ (100) orientation, with the Bi and FeO columns easily identified from the *z* contrast, as indicated in the inset atomic model. In this region, Bi atoms are displaced by about 0.35 Å away from the geometric centre of the four nearest FeO columns towards the top right. In the polarization map (Fig. 2c), the arrows represent the dipole moment, which is related to the shift of Bi atoms[14,15].

Figure 2d shows a bright-field STEM image (left) and an electric-field map (right), both derived from the scanning diffraction data in the same area as for Fig. 2b. A polarized electric field can be seen around all of the atomic columns, and a net dipole moment that points diagonally downwards is clearly revealed. A magnified image from the reconstructed HAADF is shown in Fig. 2e. The electric field in the same region is shown in Fig. 2f, where the dipole-induced polarized electric field is shown at a sampling rate of 0.2 Å per pixel. It is worth noting that the electric field surrounding the cation columns is no longer radially symmetrical: the field is weak at the top right and stronger at the bottom left across the Bi site.

Given the electric-field landscape, we can construct a 2D map of local charge density, as these two quantities are related through Gauss's law[16]:

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}$$

where ρ is the charge density and ε_0 is the vacuum permittivity. Although this equation is specified for three-dimensional (3D) geometries, we may drop the *z* dependence by integrating both sides along the *z* axis (the direction of the electron beam). Figure 3a, b shows 2D charge-density images of $SrTiO_3$ and $BiFeO_3$ projected along the (001) plane, derived from scanning diffraction experiments. The charge-density maps contain negative contributions from both core and valence electrons and positive contributions from atomic nuclei. Note that the nuclear charge distributions appear as broad Gaussians because of the shielding effect from core electrons and the size of the electron probe. For direct comparison of theory and experiment, we used Gaussian

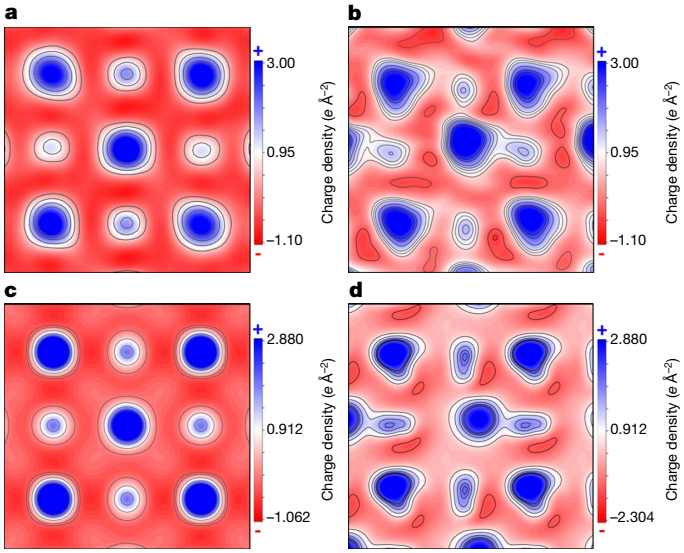


Fig. 3 | Real-space charge-density mapping in $SrTiO_3$ and $BiFeO_3$. a, Charge-density map, ρ, of bulk $SrTiO_3$ from scanning diffraction experiments. **b**, Charge-density map of bulk $BiFeO_3$. **c**, Charge-density image of bulk $SrTiO_3$ obtained from DFT calculations. **d**, Charge-density image of bulk $BiFeO_3$ obtained from DFT calculations. The simulated positive-charge distributions for the nuclear charge and the core electrons are constructed using a three-dimensional Gaussian function.

distributions to construct the positive-charge distributions of the nuclear charge and the core electrons, added them to the negative-charge distributions of the valence electrons obtained from DFT calculations, and then projected the total charge-density distribution from three to two dimensions (see Fig. 3c, d for $SrTiO_3$ and $BiFeO_3$, respectively; see Methods for the construction of the positive-charge density). The similarities in the key features shown in Fig. 3 (Fig. 3a, b from experiment and Fig. 3c, d from theory) suggest that RSCDI indeed reliably reveals the details of the charge distribution between atoms in crystalline solids.

One immediate outcome of this charge-density mapping is the clear appearance of the O atomic columns in both materials. Locating O atoms is important for studying the octahedron rotation that is often involved in the phase transitions and changes in the physical properties of complex oxides such as $La_{1-x}Sr_xMnO_3$ (ref. [17]) and Sr_2RuO_4 (ref. [18]). For $SrTiO_3$ (Fig. 3a, c), the O columns are thin, indicating no rotation of O octahedra away from the principal axis. In sharp contrast, the O columns for $BiFeO_3$ are elliptical (Fig. 3b, d) owing to a large rotation of O octahedra (by about 11°, according to DFT calculations), resulting in the splitting of the O columns in the 2D projection (Fig. 2a).

In the interstitial regions, the negative-charge density of $SrTiO_3$ (Fig. 3a) displays a four-fold symmetry around the columns of Sr atoms, matching the crystallographic symmetry of this material. Although the ionic nature of bonding in $SrTiO_3$ dominates the interatomic interaction, the existence of intense red regions suggests an accumulation of electron charge between O atoms and cations, indicating the small covalent characteristics of Sr–O and Ti–O bonds in $SrTiO_3$. In contrast, the 2D charge-density image of $BiFeO_3$ shows the deformed shapes of positive charge on all atomic sites. As shown in Fig. 3b, d, the charge contours on Fe sites exhibit triangular geometry, which originates from the O atoms shifting away from the principal axia and anisotropic FeO bonds with *d* orbitals of Fe atoms. The positive charge-density pocket of Bi shows a partial connection to the closest O column, and the existence of more intense red pockets between cations and O indicates the stronger covalent nature of BiO and FeO bonds in $BiFeO_3$ than of the bonds in $SrTiO_3$. This is also supported by analysis of the crystal orbital Hamiltonian populations in Bi-based ferroelectric materials such as $BiMnO_3$ (ref. [13]). We note that regions of intense negative-charge density migrate to the bottom left of the nuclei, from which we can directly image the electric polarization in $BiFeO_3$, along with the O octahedron tilt.

As a further step, we calculated the positions of weighted centres of positive and negative charge in $BiFeO_3$, which allowed us to determine the effective charge separation in a unit cell—a key physical quantity of multiferroic materials. We found that the separation between the positive and negative charge centres is about 0.57 Å along the diagonal axis projected in the (100) plane (Extended Data Fig. 4). According to the definition of the dipole moment, $\mathbf{p} = q\mathbf{d}$ (where q is the charge and $\mathbf{d}$ is the displacement), the measured $\mathbf{p}$ is $(3.12 \pm 0.9)e$ Å per unit cell, corresponding to a ferroelectric polarization of 78 ± 23 μC cm^{-2} along the (111) direction—comparable to the reported value of 100 μC cm^{-2} (ref. [19]). Therefore, high-resolution RSCDI enables us not only to directly see the charge polarization in $BiFeO_3$, but also to quantitatively determine its dipole moment.

The ability of RSCDI to quantify physical characteristics at high spatial resolution makes it possible to study the complex interfaces found in oxides[20,21], where many intriguing phenomena have been reported, such as the existence of superconductivity[22–24], polarization vertices[15], the quantum Hall effect[25] and magnetism[12,22,26]. Applications based on these phenomena require a better understanding of the interplay among lattices, electrons, orbitals and spin at the interface[20]. $BiFeO_3$ is multiferroic and the effect of its strong electric polarization is predicted to cause band bending in adjacent $SrTiO_3$, which, in turn, exhibits several emergent properties such as controllable interfacial conduction and photovoltaics[27]. However, owing to the absence of detailed experimental information regarding charge rearrangement in the interfacial region, understanding of the interplay among the atomic structure, charge, orbital and spin relies almost solely on DFT calculations, which are also limited by the size and complexity of the system.

Here we used RSCDI to image the electric field and charge distribution at the interface in a thin film of $BiFeO_3$ grown on (001) $SrTiO_3$. As suggested by the structural model obtained from DFT calculations (Fig. 4a), the $SrTiO_3$ substrate connects to $BiFeO_3$ by sharing a BiO layer at the interface. In Fig. 4b, the HAADF-STEM image shows a sharp contrast between weak Sr intensity and strong Bi intensity. The atomically sharp interface is also seen in atomic-resolution X-ray energy dispersive spectrum (EDS) mapping (Extended Data Fig. 5). As shown in Fig. 4c, d, we mapped the electric field and the projected charge density in the same area as in Fig. 4b. The electric dipoles are visible in $BiFeO_3$, particularly near the Bi columns. On approaching the interface, the electric dipoles become weaker in the bottom $BiFeO_3$ layers, and some induced dipoles appear around the Sr columns of the top $SrTiO_3$ layers in the substrate; a similar phenomenon involving the electric field penetrating the insulator has been proposed at $PbZr_{0.2}Ti_{0.8}O_3/SrTiO_3$ interfaces[26].

In the charge-density image, the interface effect between the insulating $SrTiO_3$ and the ferroelectric $BiFeO_3$ is revealed in three folds at the atomic scale: the O octahedron rotation, the electric polarization, and the valence charge state of Ti. As shown in Fig. 4d, the oxygen columns are well resolved in $SrTiO_3$ but become vague and elongated in $BiFeO_3$ owing to the rotation of FeO octahedra. However, the O atoms at the interface appear to have less positivity than in $SrTiO_3$ and less elongation than in $BiFeO_3$, indicating that the O octahedra are in an intermediate state. In $BiFeO_3$, the separation between the positive and negative charge pockets is clear but weakens gradually when approaching the interface. Surprisingly, the charge separation persists in $SrTiO_3$, with a smaller amplitude than in $BiFeO_3$. In Fig. 4e, f, we correlatively plot the atomic displacement, O octahedron rotation and charge separation in each unit cell (see Methods and Extended Data Fig. 6). Both atomic displacement and octahedral rotation change rapidly across the interface, with the Bi displacement of roughly 0.35 Å in $BiFeO_3$ changing to the Sr displacement of roughly 0.13 Å in $SrTiO_3$, and the octahedral rotation falling sharply from 10° to less than 1°. The charge separation is more

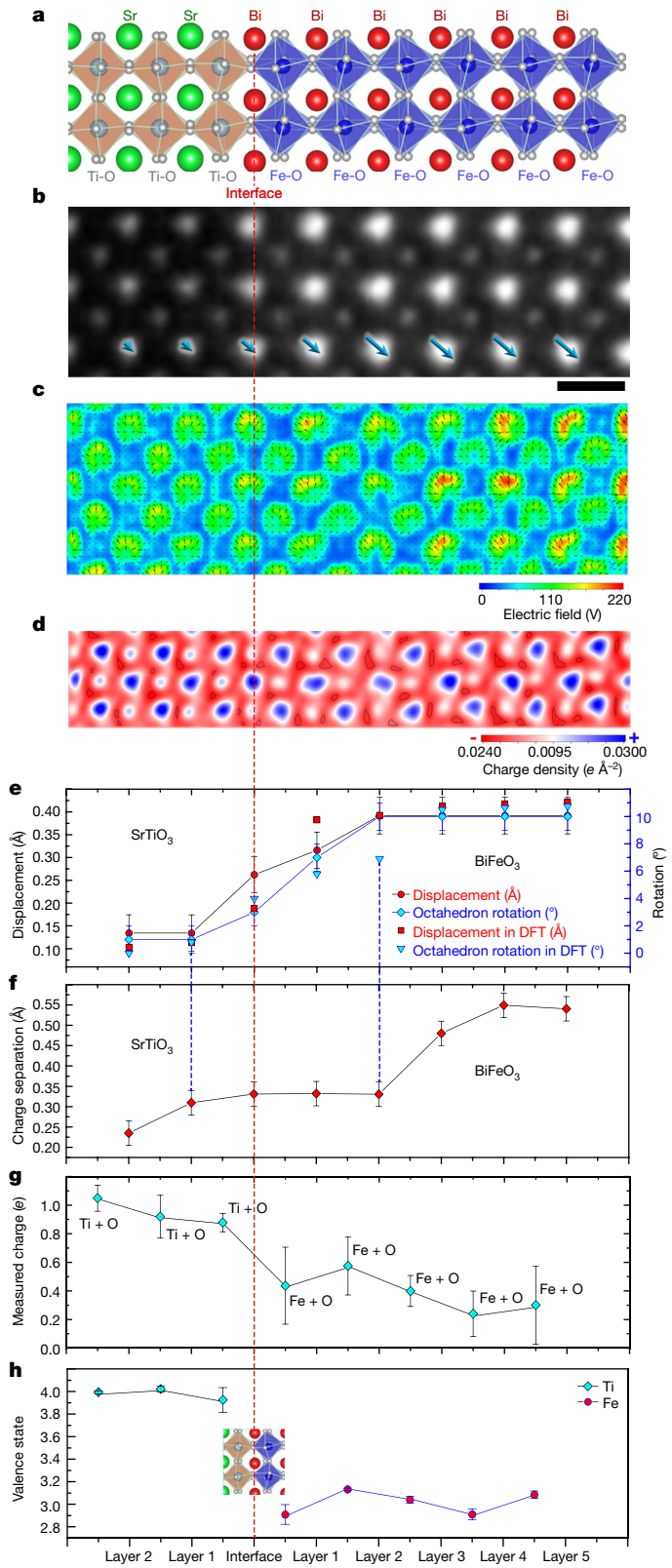


Fig. 4 | Charge-density map, O octahedron rotation and valence charge state at the interface between $SrTiO_3$ and $BiFeO_3$. a, Atomic structure of $SrTiO_3$/$BiFeO_3$ obtained from DFT calculations. **b**, Atomic-resolution HAADF-STEM image of $SrTiO_3$/$BiFeO_3$. Scale bar, 4 Å. Arrows show the direction and relative magnitude of Bi displacement. **c**, Corresponding electric-field map derived from the scanning diffraction dataset. **d**, Charge-density map for $SrTiO_3$/$BiFeO_3$. **e**, Changes in A-site displacement across the interface (that is, changes in the displacement of the Bi or Sr atom from the geometric centre of the four nearest Fe or Ti atoms), determined experimentally and by DFT calculation (in ångströms; error bars denote standard deviation; see Methods). Also shown is the O octahedron rotation determined experimentally (in degrees; error bars denote the detection limit) and from DFT calculations (scattered points). **f**, Charge separation between weighted centres of positive and negative charge within unit cells across the interface. Error bars denote the detection limit. **g**, Total charge of Ti + O and Fe + O on the two sides of the interface, measured using RSCDI. Error bars denote standard deviation. **h**, Valence states of Ti and Fe measured using high-energy-resolution EELS. Error bars denote standard deviation.

adaptive, dropping early in the second layer of $BiFeO_3$ from 0.54 Å to 0.33 Å and remaining the same across the interface until after the first layer of $SrTiO_3$, where it drops further to 0.23 Å. The atomic structure combined with the electron charge relocation shows that the electrons are most responsive to the propagation of the electric field of the polarized $BiFeO_3$, while the atomic displacement is more rigidly affected by the atomic strain. This unsynchronized response takes places in the interfacial region of three unit cells (between the two blue dashed lines in Fig. 4e, f) and leads to the phenomenon of interface charging, which is the key to understanding and engineering the 2D electron or hole gas localized at the interface.

To confirm the existence of an electron-rich interface, we calculated the total charge in regions covering the Ti + O and Fe + O columns by integrating the measured charge density (see Methods and Extended Data Figs. 7–9). Figure 4g shows the total charge of the Ti + O and Fe + O columns layer by layer; the charge of both columns drops at the interface. Figure 4h and Extended Data Fig. 10 show the results of spatially resolved electron-energy-loss spectroscopy (EELS) on Ti[28] and Fe[29,30]: the Ti valence state decreases from 4^+ to 3.7^+, indicating a mixed state of 4^+ and 3^+. The lowered Ti valence state seen in RSCDI and EELS reveals the accumulation of electrons, a phenomenon bearing similarity to the electron liquid that has been proposed at other oxide interfaces[31,32] and observed using in-line electron holography[33]. Thus, using RSCDI, we have demonstrated directly that the $SrTiO_3$/$BiFeO_3$ interface is electron rich, emerging from differences in how the atomic structure and the surrounding electron density evolve.

In summary, we have developed a new way of mapping the local charge density of materials in real space with sub-ångstrom resolution STEM and established the validity of this technique by side-by-side comparison with DFT calculations for $SrTiO_3$ and $BiFeO_3$. We have also revealed the atomic-scale charge density at the interface of $SrTiO_3$ and $BiFeO_3$, as well as variation in electric dipoles and valence states in the interfacial region. The ability to experimentally trace electron redistribution and to probe local bonding in heterogeneous materials at the subatomic level should have a substantial impact on the characterization and design of functional materials.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1649-6.

1. Midgley, P. A. Electronic bonding revealed by electron diffraction. *Science* **331**, 1528–1529 (2011).
2. Zuo, J. M., Kim, M., O'Keeffe, M. & Spence, J. C. H. Direct observation of *d*-orbital holes and Cu–Cu bonding in Cu_2O. *Nature* **401**, 49–52 (1999).
3. Nakashima, P. N. H., Smith, A. E., Etheridge, J. & Muddle, B. C. The bonding electron density in aluminum. *Science* **331**, 1583–1586 (2011).
4. Han, Z. et al. Imaging the halogen bond in self-assembled halogenbenzenes on silver. *Science* **358**, 206–210 (2017).
5. Altman, E. I., Baykara, M. Z. & Schwarz, U. D. Noncontact atomic force microscopy: an emerging tool for fundamental catalysis research. *Acc. Chem. Res.* **48**, 2640–2648 (2015).

6. Mohn, F., Gross, L., Moll, N. & Meyer, G. Imaging the charge distribution within a single molecule. *Nat. Nanotechnol.* **7**, 227–231 (2012).
7. Shibata, N. et al. Differential phase-contrast microscopy at atomic resolution. *Nat. Phys.* **8**, 611–615 (2012).
8. Shibata, N. et al. Direct visualization of local electromagnetic field structures by scanning transmission electron microscopy. *Acc. Chem. Res.* **50**, 1502–1512 (2017).
9. Müller-Caspary, K. et al. Measurement of atomic electric fields and charge densities from average momentum transfers using scanning transmission electron microscopy. *Ultramicroscopy* **178**, 62–80 (2017).
10. Müller, K. et al. Atomic electric fields revealed by a quantum mechanical approach to electron picodiffraction. *Nat. Commun.* **5**, 5653 (2014).
11. LeBeau, J. M., Findlay, S. D., Allen, L. J. & Stemmer, S. Position averaged convergent beam electron diffraction: theory and applications. *Ultramicroscopy* **110**, 118–125 (2010).
12. Wang, J. et al. Epitaxial $BiFeO_3$ multiferroic thin film heterostructures. *Science* **299**, 1719–1722 (2003).
13. Seshadri, R. & Hill, N. A. Visualizing the role of Bi 6s "lone pairs" in the off-center distortion in ferromagnetic $BiMnO_3$. *Chem. Mater.* **13**, 2892–2899 (2001).
14. Kim, Y.-M. et al. Direct observation of ferroelectric field effect and vacancy-controlled screening at the $BiFeO_3/La_xSr_{1-x}MnO_3$ interface. *Nat. Mater.* **13**, 1019–1025 (2014).
15. Nelson, C. T. et al. Spontaneous vortex nanodomain arrays at ferroelectric heterointerfaces. *Nano Lett.* **11**, 828–834 (2011).
16. Gauss, C. F. in *Werke: Fünfter Band* 3–22 (Springer, 1877).
17. Moon, E. J. et al. Spatial control of functional properties via octahedral modulations in complex oxide superlattices. *Nat. Commun.* **5**, 5710 (2014).
18. Thomas, S. et al. Localized control of Curie temperature in perovskite oxide film by capping-layer-induced octahedral distortion. *Phys. Rev. Lett.* **119**, 177203 (2017).
19. Li, J. et al. Dramatically enhanced polarization in (001), (101), and (111) $BiFeO_3$ thin films due to epitiaxial-induced transitions. *Appl. Phys. Lett.* **84**, 5261–5263 (2004).
20. Yu, P., Chu, Y.-H. & Ramesh, R. Oxide interfaces: pathways to novel phenomena. *Mater. Today* **15**, 320–327 (2012).
21. Spaldin, N. A., Cheong, S.-W. & Ramesh, R. Multiferroics: past, present, and future. *Phys. Today* **63**, 38–43 (2010).
22. Bert, J. A. et al. Direct imaging of the coexistence of ferromagnetism and superconductivity at the $LaAlO_3/SrTiO_3$ interface. *Nat. Phys.* **7**, 767–771 (2011).
23. Li, L., Richter, C., Mannhart, J. & Ashoori, R. C. Coexistence of magnetic order and two-dimensional superconductivity at $LaAlO_3/SrTiO_3$ interfaces. *Nat. Phys.* **7**, 762–766 (2011).
24. Reyren, N. et al. Superconducting interfaces between insulating oxides. *Science* **317**, 1196–1199 (2007).
25. Tsukazaki, A. et al. Quantum hall effect in polar oxide heterostructures. *Science* **315**, 1388–1391 (2007).
26. Zhang, Y. et al. Discovery of a magnetic conductive interface in $PbZr_{0.2}Ti_{0.8}O_3/SrTiO_3$ heterostructures. *Nat. Commun.* **9**, 685 (2018).
27. Guo, R. et al. Non-volatile memory based on the ferroelectric photovoltaic effect. *Nat. Commun.* **4**, 1990 (2013).
28. Ohtomo, A., Muller, D. A., Grazul, J. L. & Hwang, H. Y. Artificial charge-modulation in atomic-scale perovskite titanate superlattices. *Nature* **419**, 378–380 (2002).
29. Tan, H., Verbeeck, J., Abakumov, A. & Van Tendeloo, G. Oxidation state and chemical shift investigation in transition metal oxides by EELS. *Ultramicroscopy* **116**, 24–33 (2012).
30. Rojac, T. et al. Domain-wall conduction in ferroelectric $BiFeO_3$ controlled by accumulation of charged defects. *Nat. Mater.* **16**, 322–327 (2017).
31. Jang, H. W. et al. Metallic and insulating oxide interfaces controlled by electronic correlations. *Science* **331**, 886–889 (2011).
32. Ohtomo, A. & Hwang, H. Y. A high-mobility electron gas at the $LaAlO_3/SrTiO_3$ heterointerface. *Nature* **427**, 423–426 (2004); corrigendum **441**, 120 (2006).
33. Lee, H. et al. Direct observation of a two-dimensional hole gas at oxide interfaces. *Nat. Mater.* **17**, 231–236 (2018).

Methods

Materials and sample preparation

$BiFeO_3$ films are grown on single-crystal (100) $SrTiO_3$ surfaces by reactive molecular beam epitaxy (MBE). The 50-nm $BiFeO_3$ layer is deposited at a substrate temperature of 625 °C in distilled ozone (roughly 80% ozone) at a partial pressure of 1×10^{-6} torr. Transmission electron microscopy (TEM) samples of $SrTiO_3$ and $BiFeO_3$ thin films on $SrTiO_3$ are prepared by hand polishing followed by ion milling, to provide a large, uniformly thin area transparent to an electron beam.

Imaging methods for scanning diffraction and instrumentation

Scanning diffraction datasets are collected by synchronizing the scanning beam in STEM mode with one of the primary cameras on a microscope that is generally used to capture TEM images or diffraction patterns[10]. During STEM imaging, the electron beam is converged into an electron probe with a width of less than 0.6 Å (with aberration correction). In the detector plane, the electron beam forms a convergent beam electron diffraction (CBED) pattern. Conventional STEM images are built up by collecting the scattered or transmitted electrons during the raster of the electron probe over the sample; using high-angle annular dark field (HAADF), annular bright field (ABF) and bright-field detectors, electrons scattered into different solid angles yield a single value for each position of the scanning probe. In scanning diffraction, we collect the entire CBED pattern as a 2D image for each position; in combination with the 2D scanning area, this results in a 4D dataset.

In experiments, the 4D data are collected on a JEOL JEM 300CF double-aberration-corrected STEM at 300 kV; CBED patterns are recorded with a Gatan OneView camera at a speed of 300 frames per second (fps), each frame having a size of 512 × 512 pixels. A semi-convergence angle of 32 mrad is used for probe forming in imaging electric field. Beam scanning is synchronized with the camera using the Gatan STEMx system, with a scanning step size of 0.2 Å.

In previous attempts to derive charge-density information through scanning diffraction, noise stemmed from two main factors. First, the slow acquisition rate of the pixellated camera (30–40 fps) exacerbated the effects of sample drift, electron-beam damage and scanning noise, which all lead to image noise that obscures the charge information. At the same time, owing to the slow scanning rate, finer scanning with more sampling points cannot be achieved: finer scanning is required to show the detailed electric field and 2D charge image. Second, the insufficient signal-to-noise ratio and number of pixels for capturing the diffraction pattern during scanning will lower the signal-to-noise ratio and accuracy in measuring the electric field and charge.

In our experiment, we have improved on these factors. First, we use a Gatan OneView camera and K2 camera, which have a base acquisition rate of 300 fps; this offers both faster scanning and more sampling, minimizing sample drift during scanning and improving spatial resolution. Second, during imaging we use a weak beam intensity with a larger camera length, projecting the CBED patterns onto the 512 × 512 pixel camera. In combination with the high dynamic range of the complementary metal oxide semiconductor (CMOS)-based cameras, this enhances the details of the diffraction patterns.

Image construction

HAADF and bright-field images are reconstructed from 4D data by integrating the intensity of the annular regions of 65–172 mrad and 0–32 mrad, respectively, from the CBED patterns. These virtual detectors are analogous to the physical HAADF and bright-field detectors in STEM imaging.

Calculation of electric field

Owing to the negative charge of electrons, when an electron probe transmits through an electric field, the electrons will be deflected. In the simplified model, in a uniform electric field, the probe will shift in the diffraction plane in a negatively proportional way to the electric field, owing to the momentum change induced by the electric field. In our scanning diffraction experiment, the use of the pixellated electron detector allows the acquisition of each diffraction pattern in its entirety, as well as the calculation of the momentum change on the basis of the redistribution of diffraction intensity.

Calculation of the electric field is based on the change in momentum of the electron beam. To analyse the scanning diffraction dataset, we apply a circular mask to eliminate intensity from the high angles (more than 64 mrad), because it has been shown that scattering at high angles is less sensitive than low-angle scattering to the momentum transfer of the primary beam. In addition, by eliminating much of the detector's area where the measured intensity is low, we eliminate a large source of noise from the centre-of-mass calculation. We have found that this mask substantially decreases the noise level in our final electric-field images.

Next, we calculate the COM of each diffraction pattern in the dataset; the centre of the diffraction pattern is determined by averaging the COM within 64 mrad of all diffraction patterns. Then, at each scanning point, the deviation of the COM from the centre of the diffraction pattern gives us a vector field. As shown in ref. [9], the COM of the diffraction pattern is equivalent to the momentum transfer. Extended Data Fig. 1 shows the strength of the electric-field change at different locations within $SrTiO_3$ samples of varying thicknesses, with the scanning diffraction data being generated using multi-slice simulation[34]. For samples thinner than 6 nm, the momentum transfer is proportional to the increase in sample thickness. Fitting the electric-field strength versus thickness plot with a linear curve shows that the precision measured by standard deviation is 3.5% for weak field regions between atoms (blue), 1.3% for areas closer to atoms (red), and 6.7% for areas very close to atoms (black). The change in the linear relationship is due to the beam being continuously shifted when propagating through the sample. For thicker samples, the accuracy gradually falls off. In samples with thicknesses of around 6 nm, the centre of weight can be still interpreted as being representative of the local electric field, with some loss in quantitative accuracy.

To measure the electric field in detail, the procedure is as follows. First, calibrate the pixel size in momentum space (in units of mrad per pixel). Second, the momentum change of electrons is based on:

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = -e\mathbf{E}$$

Solving this differential equation with the assumptions outlined in ref. [9], the momentum change can be written in terms of a few simple parameters:

$$\Delta\mathbf{p}_{xy} = -\left(\frac{e\Delta z}{v_z}\right)\mathbf{E}_{xy}$$

Where $\Delta\mathbf{p}_{xy}$ is the momentum change, e is the charge of an electron, Δz is the sample thickness, v_z is the speed of the electron along the beam direction and $\mathbf{E}_{xy}$ is the electric field. Given that the diffraction pattern is a momentum space image of the probe, $\Delta\mathbf{p}_{xy}$ can be calculated from the shift in the COM of the diffraction pattern:

$$\Delta\mathbf{p}_{xy} = \Delta\mathbf{COM}p_z$$

where $\Delta\mathbf{COM}$ is the shift in the COM from the geometric centre (in milliradians) and p_z is the momentum of the electron beam along the beam direction. Third, the electric-field strength can be calculated using:

$$\mathbf{E}_{xy} = -\frac{\Delta\mathbf{p}_{xy}v_z}{e\Delta z} = -\frac{\Delta\mathbf{COM}p_z v_z}{e\Delta z}$$

Measurement of specimen thickness

Position-averaged CBED (PACBED) is carried out by acquiring the average CBED pattern in the diffraction plane while the electron beam scans across a small area. A smaller convergence angle of approximately 10.6 mrad is used. Previous work[11] has shown that the thickness and orientation of the sample have a large influence on the PACBED pattern. Here, by recording CBED patterns with a low convergence angle and incoherently averaging over larger regions (roughly 1 nm^2), we compare the experimental results with simulated PACBED patterns to quantitatively determine the thickness of our samples. The database of simulated PACBED patterns is generated using multi-slice simulation[34] containing PACBED patterns from different sample thicknesses at the same imaging condition (300 kV, 10.6 mrad). The sample thickness is then estimated at the precision of around 1 nm. The tilt of the sample is also more apparent under these conditions than in standard imaging conditions, so we use PACBED to align the sample to the zone axis more accurately during data acquisition.

The PACBED in Extended Data Fig. 2 is taken experimentally from the boxed area of $SrTiO_3$ labelled in the HAADF-STEM image in Extended Data Fig. 2a. This PACBED is compared with simulated PACBEDs from samples of thickness 0.8–16 nm. Quantitatively, least-squares fitting is carried out on PACBEDs from an $SrTiO_3$ sample and from the $SrTiO_3$ part of the interface, showing that the $SrTiO_3$ region is around 5.6 ± 1 nm thick. Extended Data Fig. 3 shows a PACBED taken from the selected $BiFeO_3$ region; this PACBED is comparable to simulated PACBEDs from $BiFeO_3$ samples of thickness 6–8 nm.

Mapping charge density

After deriving the electric-field landscape, we can calculate the divergence of our measured electric field and then determine the charge density by using Gauss's law: the divergence of the electric field is proportional to the charge density. After mapping the charge density in a unit cell, we can separate the positive and negative charges and calculate their weighted centres, as in Extended Data Fig. 4 for $BiFeO_3$. In $BiFeO_3$, the measured positive and negative charge centres separate along the diagonal direction with a spacing of 0.57 Å.

Image correction

Even though the acquisition speeds of fast CMOS cameras have improved substantially, reaching only a few milliseconds per frame, the dwell pixel time is still more than a hundred times slower than that of ADF and bright-field detectors at the microsecond level. Sample drift and image distortion therefore exist that can be seen from the reconstructed HAADF and bright-field images. This geometric image distortion needs to be corrected. We used 2D Gaussian fitting on the reconstructed HAADF image to determine the atom column position in the scanning diffraction experiment. Using a conventional HAADF image as a reference, the deviation of the atomic locations in scanning diffraction is compensated using geometric transformation, where each pixel location is corrected on the basis of the local correction matrix, considering lattice rotation and scaling. The corrected data with reduced distortion can then reveal more details on the electric-field distribution. Unit cells with similar patterns can be averaged to improve the signal-to-noise ratio.

In detail, the sample drift along the horizontal direction is corrected by measuring the average displacement per line from a conventional fast-scanned STEM image and then shifting each line of pixels back. Drift along the vertical scan direction is corrected by rescaling the vertical axis to fit with the fast-scanned STEM image. These operations are first performed on the reconstructed HAADF and bright-field images on the basis of the visible lattice, and the same shift and rescaling operations are then applied to the electric-field maps.

DFT calculation

First-principles DFT calculations are performed with the Vienna ab-initio simulation package (VASP)[35,36]. The spin-polarized generalized gradient approximation (GGA)[37] is used to describe the exchange–correlation interaction among electrons. We treated Bi 6*s* 6*p*, Fe 3*d* 4*s*, Sr 4*s* 4*p* 5*s*, Ti 3*d* 4*s* and O 2*s* 2*p* as valence states and adopted projector-augmented wave (PAW) pseudopotentials to represent the effects of their ionic cores[38,39]. Spin–orbit coupling is not included in the calculation. To describe the correlation effect properly, we use the GGA + *U* method for the localized *d* orbitals of Fe (Coulomb repulsion $U = 3.0$ eV; on-site exchange $J = 0.0$ eV)[40]. Calculations of $2 \times 2 \times 2$ supercells are carried out to simulate, first, the rotation of adjacent O octahedron and distorted Bi and Fe sites along the [111] direction; and second, the G-type antiferromagnetic exchange coupling between Fe atoms. We sample the Brillouin zone by adopting the Γ-centred Monkhorst–Pack method[41] with a density of about $2\pi \times 0.03$ $Å^{-1}$ in all calculations[42]. Brillouin zone integrations are performed with a Gaussian broadening of 0.05 eV during all calculations. The energy cut-off for the plane-wave expansion is 500 eV, which results in good convergence of the computed ground-state properties, according to our previous investigations of oxides[43–45]. Structures are optimized with the criterion that the atomic force on each atom becomes weaker than 0.01 eV $Å^{-1}$ and the energy convergence is better than 10^{-6} eV.

To corroborate our experimental observations and to validate our conclusion that the observed image contrast results from local charge density, we derive the electron charge density of bulk $SrTiO_3$ and $BiFeO_3$, projected along the (001) direction, through DFT calculations. It is well known that DFT allows the solving of one-electron Kohn–Sham equations and the evaluation of electron density by using a lattice potential acting on the system's electrons[46]. To appreciate the positive-charge effect from the nucleus in experiments, a Gaussian distribution of effective core charge at different nucleus sites is constructed, with the conservation condition that the positive charge is equal to the negative charge (the total number of electrons) in the system. The Gaussian broadening widths for Sr, Ti, Bi, Fe and O are 0.40 Å, 0.31 Å, 0.50 Å, 0.33 Å and 0.30 Å, respectively[47]. The DFT-calculated charge-density image is then derived by summing the electron charge density of the ground states and the Gaussian-distributed positive-charge density.

Quantification of O octahedron rotation

In RSCDI images, O can be seen from the positive-charge intensity. Without O octahedron rotation, in $SrTiO_3$, the O atoms in the face centres overlap along the projection of (100); with rotation, in $BiFeO_3$, the O atoms in face centres split along the projection. Such splitting leads to weakening and elongation of the O charge intensity.

In Extended Data Fig. 6, we plot the intensity and width from O columns along with the rotation angle in a charge-density image from DFT. In general, with higher octahedron rotation, the O intensity elongates and the peak intensity decreases. From one to seven degrees of rotation, no obvious change in the width of the oxygen intensity is seen. However, the peak intensity is more sensitive to octahedron rotation in this range, as shown by the blue points and linear fit in the plot. The linear fitting between O intensity and octahedron rotation is used to quantify the O rotation in Fig. 4e; the fitting shows that the deviations of all points are less than one degree from the linear curve, so the error bars for measured octahedron rotation are chosen to be one degree.

Measurement of total charge

The total charge from individual atomic columns is measured by integrating the intensity within a region around each atomic column in the charge-density image. To define the centre of all heavy atomic columns, we use the peak position of atoms in the HAADF image calculated by 2D Gaussian fitting. To determine the size of the inclusion area, we apply a method from DFT for calculating the charge state of atoms, Bader charge analysis[48,49], which uses the saddle points of the charge-density contour between atoms to define the boundary for electrons belonging to different atoms. As shown in Extended Data Fig. 7, in our images, we look for the local minimum in the charge density and use

this point as the boundary in calculating the charge state from the charge-density image.

In our measurements of total charge for each atomic column, note that in $SrTiO_3$, along the (100) projection, Ti and O atoms overlap, so the integrated charge is from Sr, O and Ti + O columns. In $BiFeO_3$, Fe and O overlap, so that the measured charge is from Bi and Fe + O columns. We do not consider the O columns in $BiFeO_3$ because these columns split. For each type of atomic column, the intensity is measured and shown in the histograms in Extended Data Fig. 8. From the fitted peaks, Bi intensity centres on 9.58, Sr on 5.71, Ti + O on 4.45, Fe + O on 2.51 and O on −1.50. As the measured intensity resembles the total charge, including positive nuclei and core orbital electrons, the charge intensity reflects the valence of each type of atomic column.

To estimate the partial charge of the atoms, we carry out a Bader charge calculation[48,49] for bulk $SrTiO_3$ and $BiFeO_3$. For bulk $SrTiO_3$, we find partial charges of 8.421*e*, 1.445*e* and 7.378*e* for Sr, Ti and O atoms, respectively (Extended Data Fig. 8b). Because there are respectively 10, 4 and 6 valence electrons for Sr ($4s^2\,4p^6\,5s^2$), Ti ($3d^2\,4s^2$) and O ($2s^2\,2p^4$) atoms, their corresponding valence states are $Sr^{+1.58}$, $Ti^{+2.55}$ and $O^{-1.38}$. Similarly, in bulk $BiFeO_3$, we find 2.131*e*, 6.119*e* and 7.584*e* for Bi, Fe and O atoms, respectively. Considering that there are respectively 5, 8 and 6 valence electrons for Bi ($6s^2\,6p^3$), Fe ($3d^6\,4s^2$) and O ($2s^2\,2p^4$) atoms, their corresponding valence states are $Bi^{+2.87}$, $Fe^{+1.88}$ and $O^{-1.58}$.

We note that Bader charge calculation uses saddle points (or 'zero flux surfaces') at which the charge density is a minimum in order to separate atoms from each other[48,49]. Owing to the covalent bond nature of the $SrTiO_3$ and $BiFeO_3$ systems, the charge states obtained through Bader charge analysis are usually underestimated (which is not fully consistent with the definition of 'valence' in chemistry) compared with the experimental results obtained through EELS.

In stoichiometric phase $BiFeO_3$, the valence states of Bi, Fe and O are 3^+, 3^+ and 2^-, respectively, which gives the Fe + O column a valence state of 1^+ when projected along (100). In $SrTiO_3$, the valence states of Sr, Ti and O are 2^+, 4^+ and 2^-, respectively, making the projected Ti + O column 2^+. If we order each atomic column according to its valence state (Bi, 3^+; Sr, 2^+, Ti + O, 2^+; Fe + O, 1^+; and O, 2^-), we can see that this matches the ordering of the integrated intensities found in the charge-density image. However, as discussed above, the partial charge derived from charge density—both experimentally and through DFT—could be different from the definition of 'valence' in chemistry. Therefore, we compare the total charge intensity with the partial charge derived through DFT. Now, Bi is 2.87^+, Sr is 1.58^+, Ti + O is 1.17^+, Fe + O is 0.3^+ and O is 1.38^- (Extended Data Fig. 9a). All charges are well fitted with a linear curve plotting the charge intensity against the partial charge from DFT. We then use this linear relationship to estimate the charge of Ti + O and Fe + O columns across the interface in Fig. 4g (where the error bar denotes the standard deviation from all atomic columns in a row parallel to the interface). We emphasize that the charge variation measured using RSCDI assumes a uniform sample thickness, which can be determined by PACBED.

In Extended Data Fig. 9a, the errors may derive from uncertainties in experiments and analysis, including the Poisson noise from the pixellated detector, sample drift and scanning noise in STEM, the distribution of charge intensities of each species, and uncertainties in determining the area used to calculate the integrated charge.

By comparing experimentally measured partial charges with those from DFT calculations, we show that quantification of partial charge using RSCDI can be reliable. Relative changes in atomic charge can be easily observed by comparing the partial charge of atomic columns, when the area is uniform in thickness. Fully quantitative measurement depends on the fitting curve, which can be achieved in two ways: first, by precisely measuring the sample thickness, calibrating the instrument, and using a standard sample as a reference; and second, by comparison with DFT calculations.

We also note that, when measuring the total charge, RSCDI will work best when the sample is along a zone axis in which each atomic column comprises only one atom species, so that atoms with different charge states will not overlap. The projection of 3D charge to two dimensions can also result in underestimation of those bond charges that are distributed among atoms at different heights. This effect needs further study.

Electron energy loss spectroscopy

High spatial resolution STEM-EELS experiments are carried out on a Nion UltraSTEM 200, equipped with a C_3/C_5 corrector and a high-energy-resolution monochromated EELS system (HERMES). The instrument is operated at 60 kV with a convergence semi-angle of 30 mrad and a beam current of about 100 pA. For spectrum image acquisition, we use a dispersion of 0.19 eV per channel, a dwell time of 0.4 s per pixel and a pixel size of 0.1 nm. The background in each spectrum is removed by fitting a power-law function to the pre-edge region using the commercial software package DigitalMicrograph. To reveal the valence state of Ti, we use the multiple least squares (MLS) methods to separate the Ti^{4+} and Ti^{3+} components of each Ti EEL spectrum using the following equation: $S(E) = a_1R_1(E) + a_2R_2(E) + \chi(E)$, where E is the energy loss; $S(E)$ is the experimental EEL spectrum; $R_1(E)$ and $R_2(E)$ are standard Ti^{4+} and Ti^{3+} spectra[28]; a_1 and a_2 are fit coefficients (spectral weight) of Ti^{4+} and Ti^{3+} components; and $\chi(E)$ is the residual spectrum. Fitting process is carried out using the commercial software Digital Micrograph. The valence state of Ti is further calculated from the weighted arithmetic mean of the Ti^{4+} and Ti^{3+} components (Fig. 4h). To analyse the valence state of Fe, we measure the difference between the onset energies of the O *K*-edge and Fe L_3-edge, which were determined by the energy loss when the edge reaches 10% of its maximum intensity. Owing to the linear relationship between this Fe–O onset energy difference and the valence state of Fe[29,30], we can further calculate the localized valence state of Fe (Fig. 4h; the error bar denotes the standard deviation from all points collected within the atomic columns).

There is a slight discrepancy between the charge states measured by EELS and by charge-density imaging: the charge in $SrTiO_3$ drops earlier in Fig. 4g and later in Fig. 4h. This discrepancy derives from the delocalization of the EEL spectrum, such that the signal can be influenced by adjacent regions. As measured in EELS, the slightly higher valence state of Ti close to the interface is the result of the valence of Ti from the bulk part of the $SrTiO_3$ being delocalized and detected by the probe at the interface. However, charge-density imaging is a more local probe: it is affected only by the field within its interaction volume, and we only integrate the charge density close to the atomic column. Therefore, the local charge measured by charge-density imaging more clearly shows a change in the charge state of individual atomic columns, rather than in the material as a whole, which is measured with EELS.

Data availability

The datasets generated or analysed here are available from the corresponding authors on reasonable request.

34. Zuo, J. M. & Spence, J. C. H. *Electron Microdiffraction* (Plenum, 1991).
35. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
36. Kresse, G. & Hafner, J. Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Phys. Rev. B* **49**, 14251–14269 (1994).
37. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
38. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
39. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
40. Dudarev, S. L., Botton, G. A., Savrasov, S. Y., Humphreys, C. J. & Sutton, A. P. Electron-energy-loss spectra and the structural stability of nickel oxide: an LSDA+U study. *Phys. Rev. B* **57**, 1505–1509 (1998).
41. Monkhorst, H. J. & Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **13**, 5188–5192 (1976).

42. Elsässer, C., Fähnle, M., Chan, C. T. & Ho, K. M. Density-functional energies and forces with Gaussian-broadened fractional occupations. *Phys. Rev. B* **49**, 13975–13978 (1994).
43. Wang, Z., Wang, H., Yu, C. C. & Wu, R. Q. Hydrogen as a source of flux noise in SQUIDs. *Phys. Rev. B* **98**, 020403 (2018).
44. Wang, H. et al. Candidate source of flux noise in SQUIDs: adsorbed oxygen molecules. *Phys. Rev. Lett*. **115**, 077002 (2015).
45. Wang, H. et al. Possibility of realizing quantum spin Hall effect at room temperature in stanene / Al_2O_3 (0001). *Phys. Rev. B* **94**, 035112 (2016).
46. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev*. **140**, A1133–A1138 (1965).
47. Shannon, R. D. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallogr. A* **32**, 751–767 (1976).
48. Henkelman, G., Arnaldsson, A. & Jónsson, H. A fast and robust algorithm for Bader decomposition of charge density. *Comput. Mater. Sci*. **36**, 354–360 (2006).
49. Bader, R. F. W. *Atoms in Molecules—A Quantum Theory* (Oxford Univ. Press, 1990).

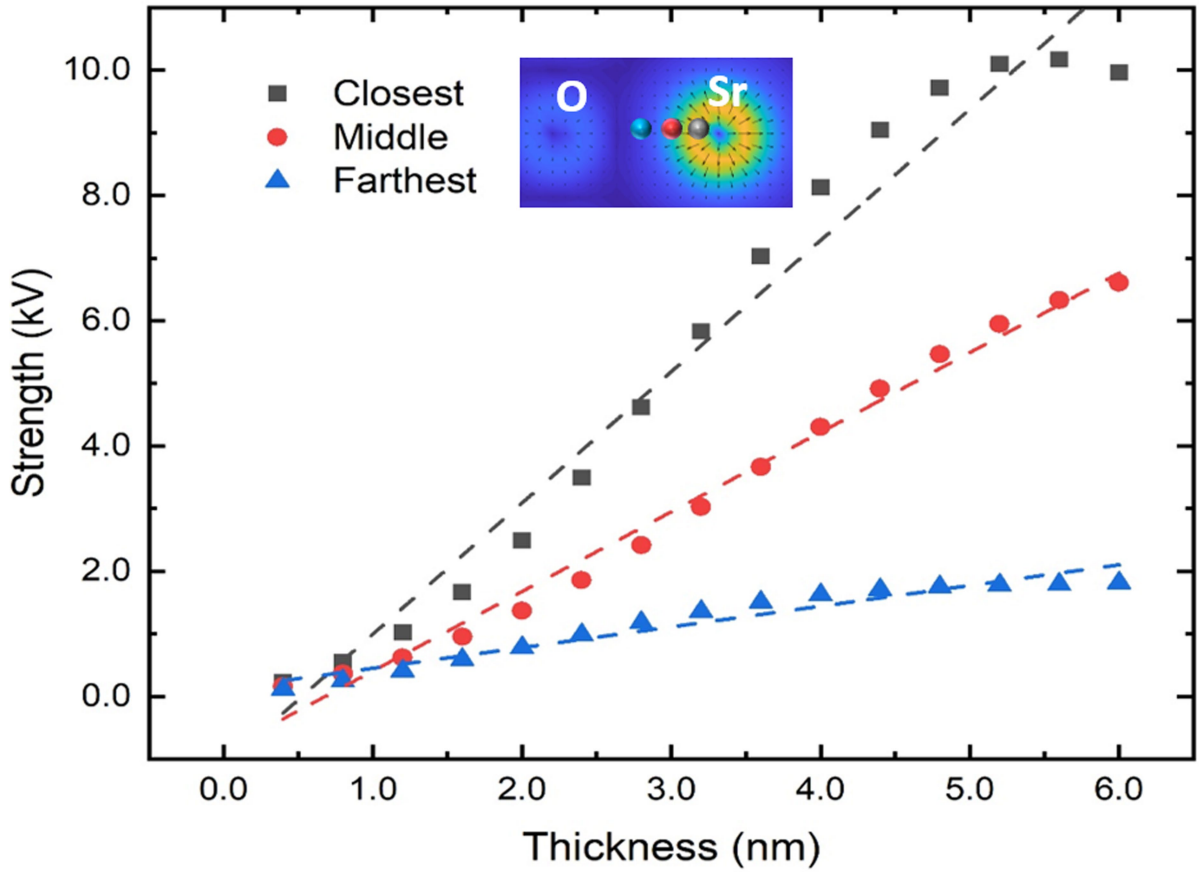
Acknowledgements Our experimental work was supported by the US Department of Energy (DOE), Office of Basic Energy Sciences, Division of Materials Sciences and Engineering under grant DE-SC0014430. TEM specimen preparation and sample thickness fitting were partially supported by the US National Science Foundation (NSF) under grant number DMR-1506535. DFT studies were supported by the US DOE (grant number DE-FG02-05ER46237) and the National Energy Research Scientific Computing Center (NERSC). Growth of $BiFeO_3$ films at Cornell University was supported by the National Science Foundation (Nanosystems Engineering Research Center for Translational Applications of Nanoscale Multiferroic Systems) under grant number EEC-1160504, and film growth at Nanjing University was supported by the National Basic Research Program of China (grant number 2015CB654901). TEM experiments were conducted using the facilities in the Irvine Materials Research Institute (IMRI) at the University of California at Irvine. We thank H. Sawada from Jeol Ltd. for help with experiments.

Author contributions W.G., R.W. and X.Q.P. conceived this project and designed the studies; W.G. and C.A. performed electron microscopy experiments and data analysis with the help of T.A.; H.W. and Y.H. carried out DFT calculations; X.Y. performed EELS experiments and analysis; Y.Z., L.L., H.H., T.B., W.G. and C.A. prepared TEM samples; D.J., C.H., Y.N. and D.S. made thin films; W.G., C.A., H.W., R.W. and X.Q.P. wrote the paper with the contributions from all authors.

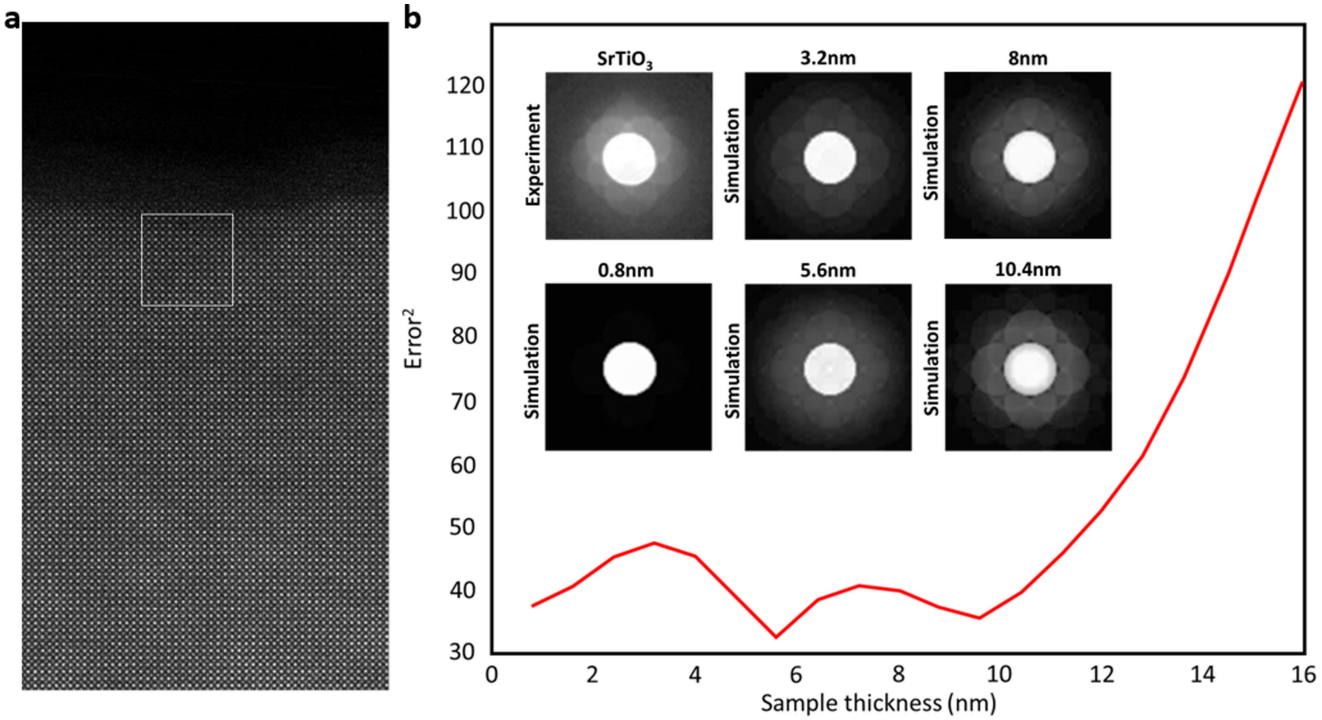
Competing interests The authors declare no competing interests.

Additional information

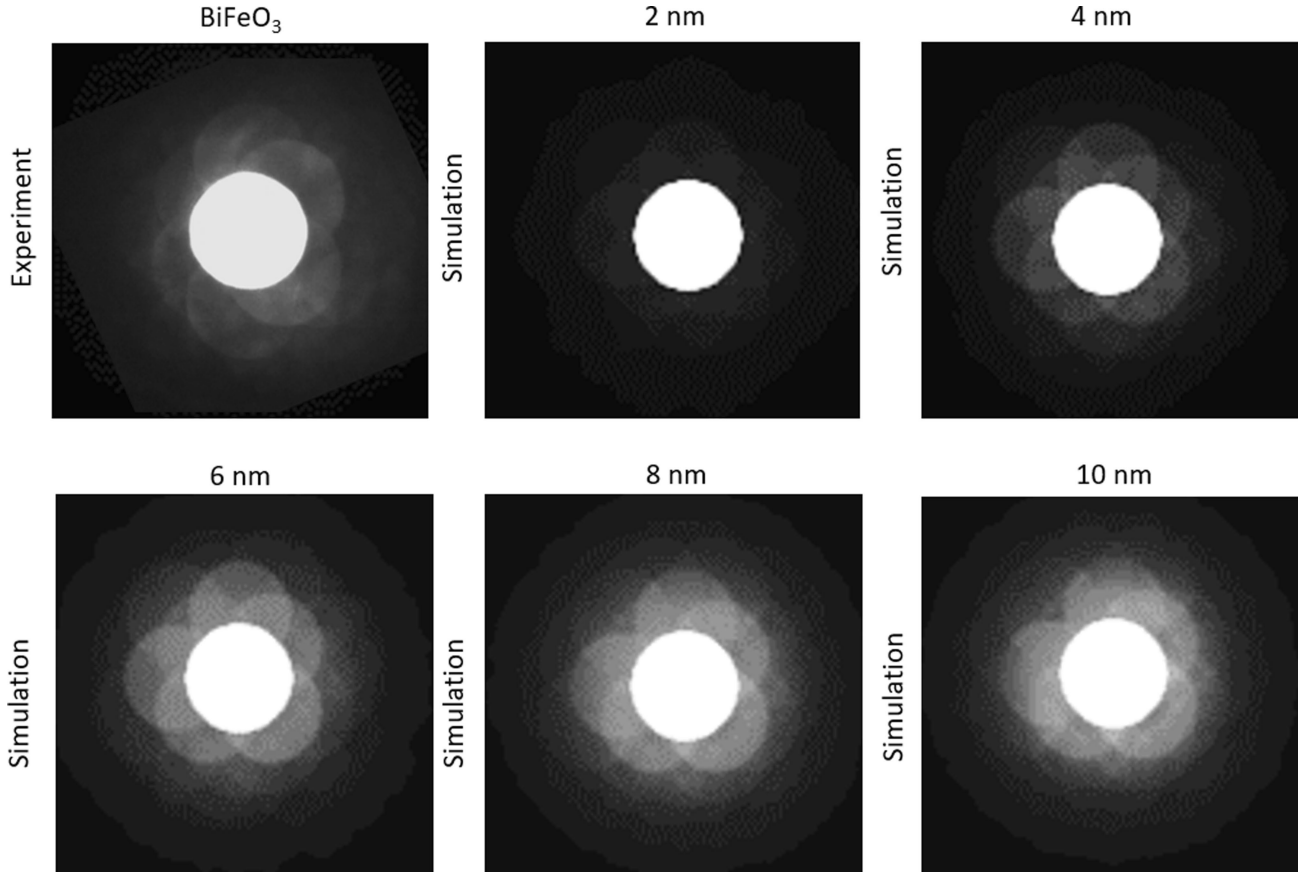
Correspondence and requests for materials should be addressed to R.W. or X.P.



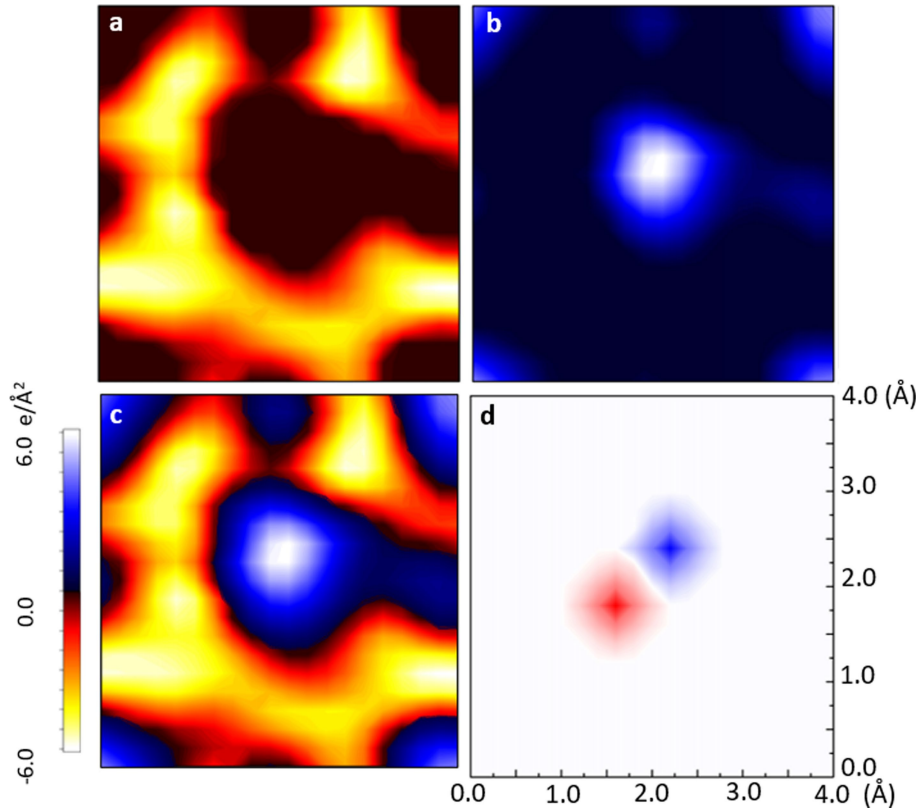
Extended Data Fig. 1 | Measured electric-field strength in $SrTiO_3$ films of different thicknesses. The electric-field strength at locations close to Sr atoms (black), farther away (red) and farthest away, in between the Sr and O atomic columns (blue), was calculated from simulated diffraction data with different sample thicknesses up to 6 nm. The measured electric-field strength is shown as points, and the dashed lines denote linear fitting. The inset shows the sampling locations for each line on a map of the simulated electric field. Diffraction data were generated using multi-slice simulations in which conditions were matched to experimental conditions.



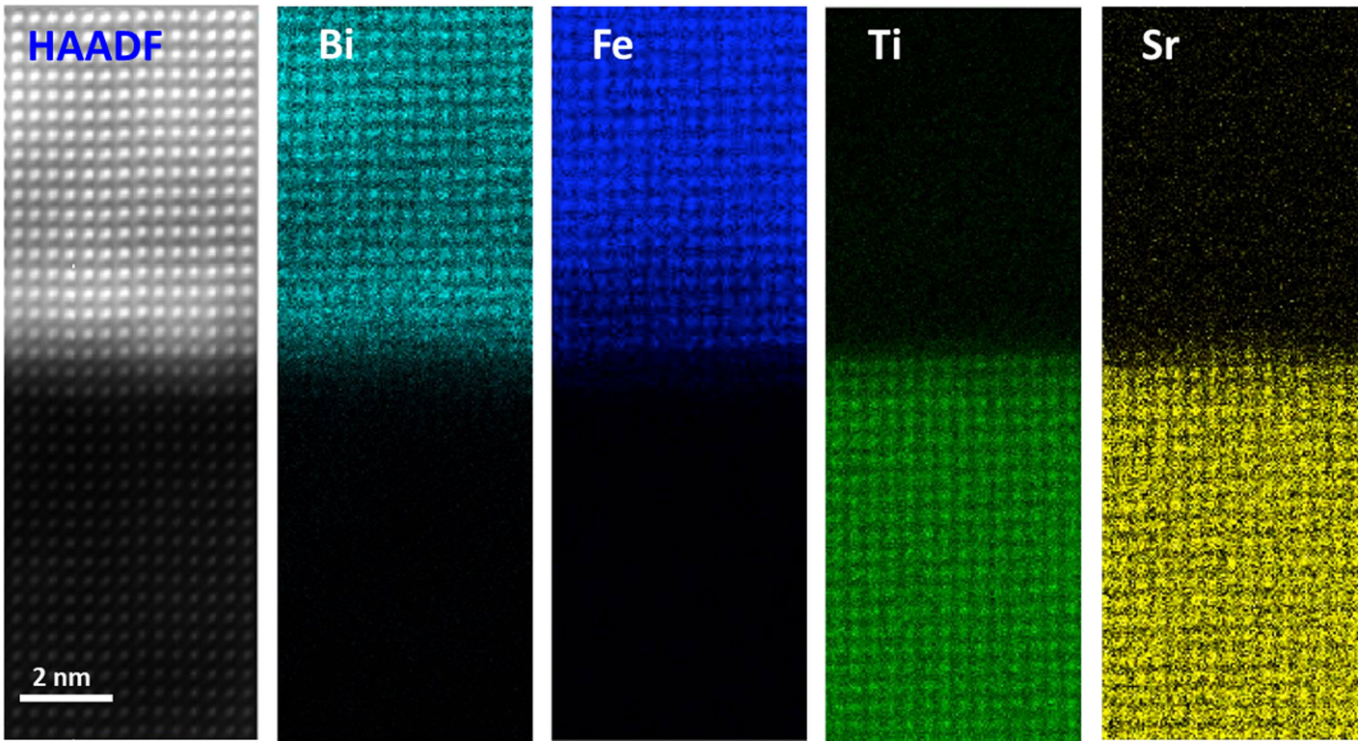
Extended Data Fig. 2 | Measuring $SrTiO_3$ thickness using PACBED. a, HAADF-STEM image of $SrTiO_3$. **b**, Least-squares fitting of the experimental PACBED results with the simulated PACBEDs (red line). The inset shows the PACBED acquired from the boxed region in the STEM image, and simulated PACBEDs of $SrTiO_3$ with thicknesses from 0.8 nm to 10.4 nm.



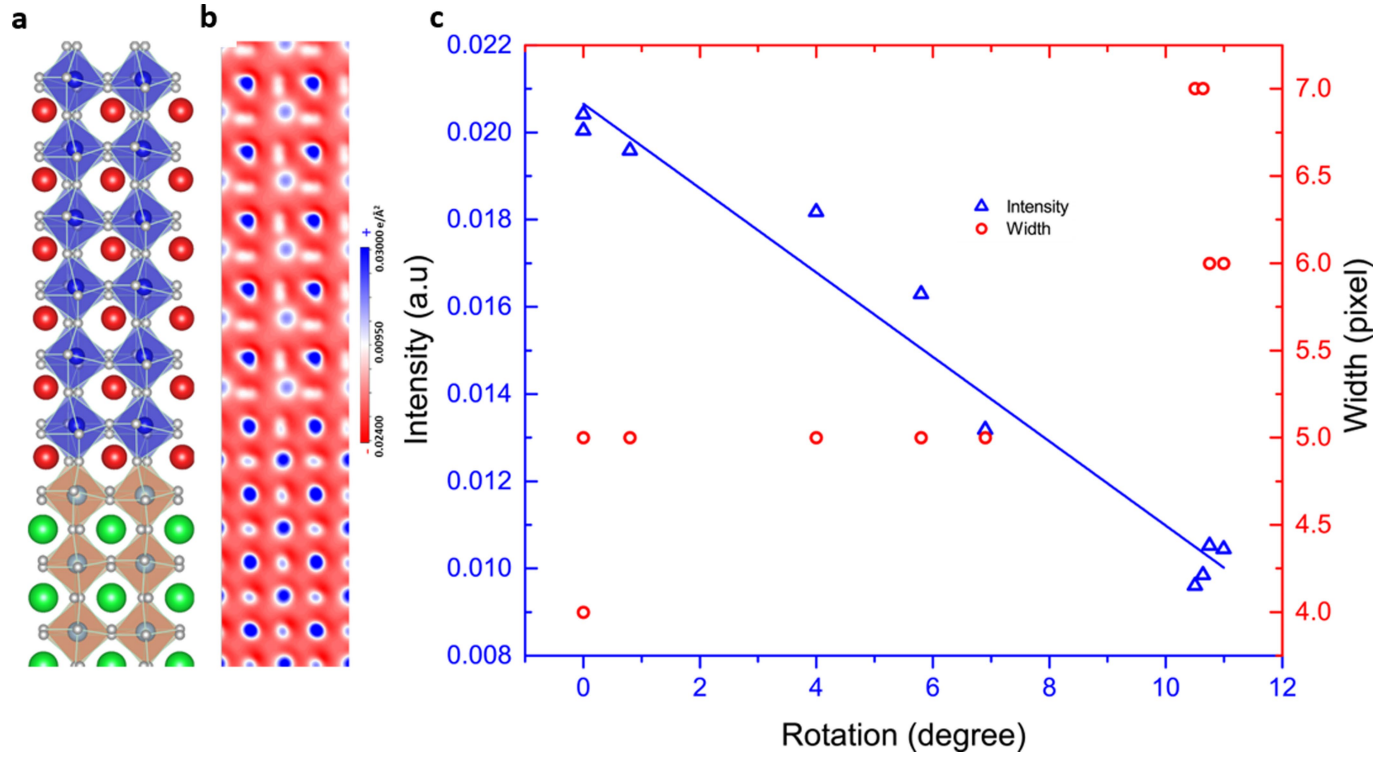
Extended Data Fig. 3 | Measuring $BiFeO_3$ thickness using PACBED. Shown are the PACBED acquired in experiment and simulated PACBEDs for $BiFeO_3$ with thicknesses of 2–10 nm.



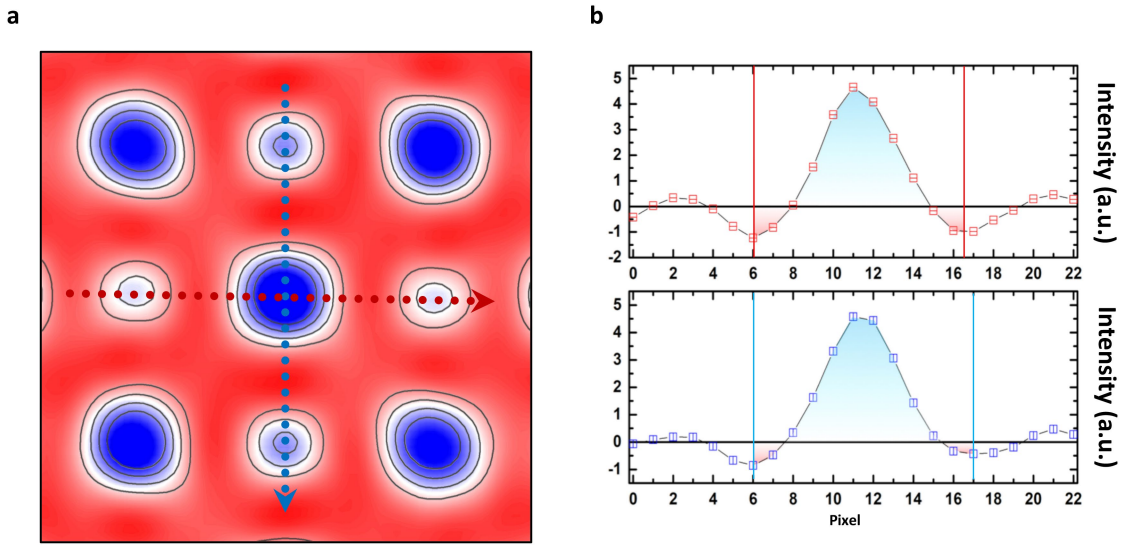
Extended Data Fig. 4 | Separation of positive and negative charge in a $BiFeO_3$ unit cell. a–**c**, Negative charge (**a**); positive charge (**b**); and overlapping of positive and negative charge (**c**) in the pseudo-cubic unit cell of $BiFeO_3$. **d**, Positions of positive (blue) and negative (red) charge centres.



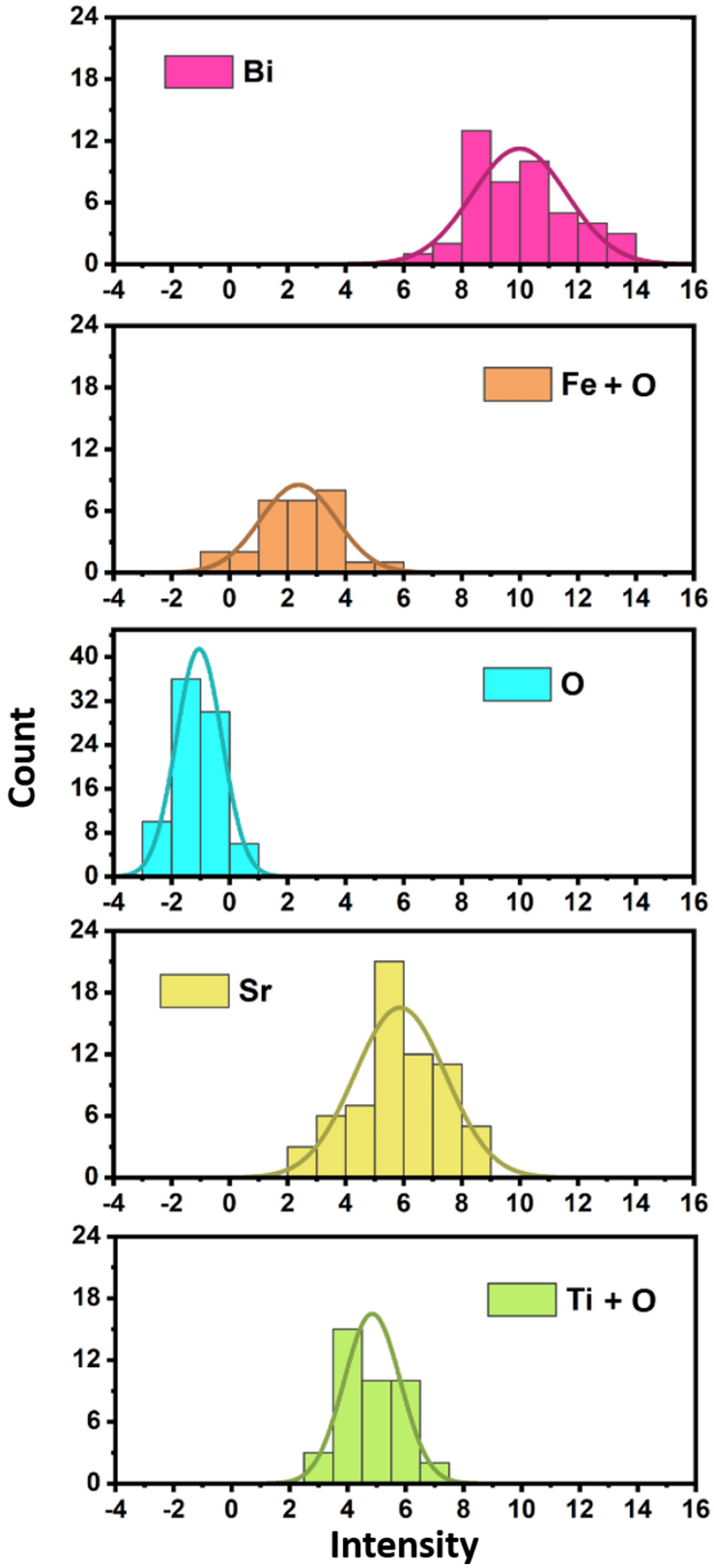
Extended Data Fig. 5 | Atomic-resolution EDS maps across the $BiFeO_3/SrTiO_3$ interface. The EDS map was acquired using a JEM300CF AC-STEM system with EDS dual silicon-drift detectors (SDDs). Thirty scans (each with a 0.4-ms dwell time) in the same area across the interface were aligned and summed. The HAADF-STEM image and atomic-resolution EDS maps of Bi, Fe, Sr and Ti reveal an atomically sharp interface.



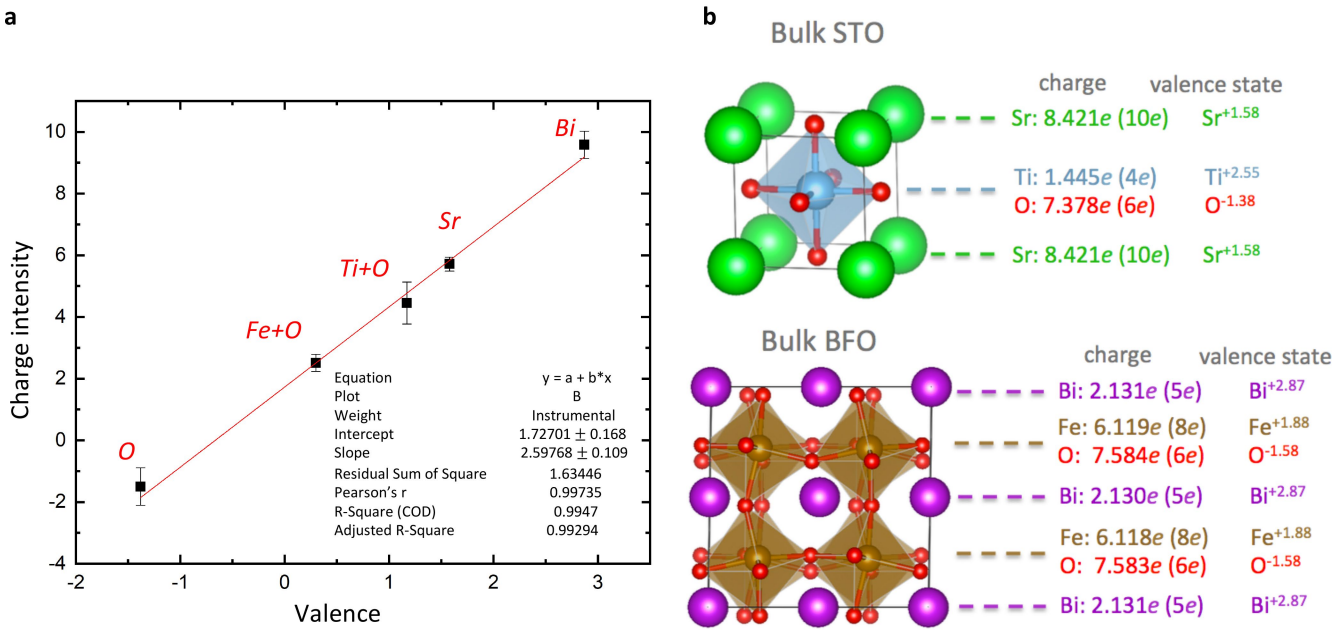
Extended Data Fig. 6 | Measurement of O octahedron rotation. a, Atomic model of the $BiFeO_3/SrTiO_3$ interface, which is relaxed and then calculated by DFT. The rotation of O octahedra is readily visible from the splitting of the O atoms in this projection. **b**, Charge-density image calculated using DFT. The images of O charge become elongated and weak with higher O octahedron rotation. **c**, Intensity of O column charge (blue) and width of O intensity (red) plotted against O octahedron rotation measured using the atomic model from DFT calculations.



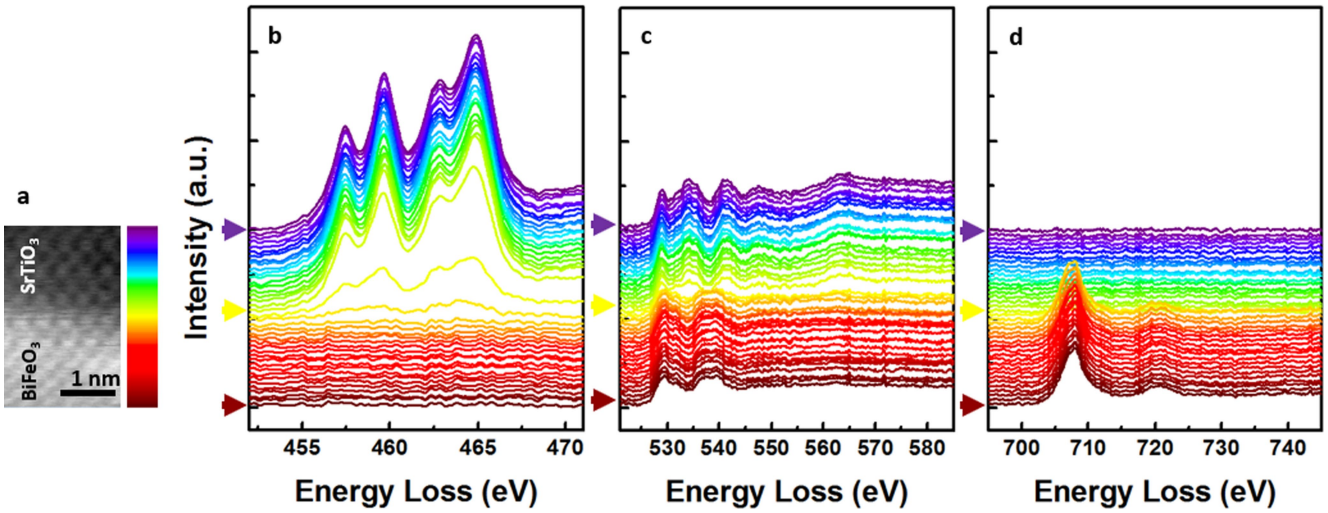
Extended Data Fig. 7 | Determination of the region for measuring the total charge of atomic columns. a, 2D charge-density image of $SrTiO_3$. **b**, Charge-intensity profile drawn along the horizontal (red) and vertical (blue) directions as shown in **a**. Local minima in the charge-intensity profile are defined as the boundary of the area included for integrating the charge.



Extended Data Fig. 8 | Measurement of the total charge of each atomic site. Histograms showing the integrated intensity of Bi columns, Fe + O columns, O columns, Sr columns and Ti + O columns from charge-density images of $BiFeO_3$ and $SrTiO_3$.



Extended Data Fig. 9 | Charge-intensity change as a function of valence. **a**, Integrated intensity in each atomic column in the charge-density image, plotted as a function of valence derived through DFT to show their correlation. The red line is the linear fit. **b**, Partial charge and valence states of all atoms derived through Bader charge analysis in DFT.



Extended Data Fig. 10 | High-resolution core-loss EELS measurement of Ti, O and Fe at the $SrTiO_3/BiFeO_3$ interface. a, HAADF-STEM image used for acquiring EELS data on the $SrTiO_3/BiFeO_3$ interface. Scale bar, 1 nm. **b**–**d**, Stacking EEL spectra of the Ti $L_{2,3}$-edge (**b**); O *K*-edge (**c**); and Fe $L_{2,3}$-edge (**d**) across the interface. The location of each coloured spectrum is marked by the colour bar in **a**. Each spectrum is averaged in the direction parallel with the $SrTiO_3/BiFeO_3$ interface. The purple, yellow and maroon arrows indicate respectively the top edge, interface and bottom edge of the mapping region in **a**.

Geochemical evidence for high volatile fluxes from the mantle at the end of the Archaean

https://doi.org/10.1038/s41586-019-1745-7

Received: 30 May 2019

Accepted: 17 September 2019

Published online: 20 November 2019

Bernard Marty[1]*, David V. Bekaert[1], Michael W. Broadley[1] & Claude Jaupart[2]

The exchange of volatile species—water, carbon dioxide, nitrogen and halogens—between the mantle and the surface of the Earth has been a key driver of environmental changes throughout Earth's history. Degassing of the mantle requires partial melting and is therefore linked to mantle convection, whose regime and vigour in the Earth's distant past remain poorly constrained[1,2]. Here we present direct geochemical constraints on the flux of volatiles from the mantle. Atmospheric xenon has a monoisotopic excess of ^{129}Xe, produced by the decay of extinct ^{129}I. This excess was mainly acquired during Earth's formation and early evolution[3], but mantle degassing has also contributed ^{129}Xe to the atmosphere through geological time. Atmospheric xenon trapped in samples from the Archaean eon shows a slight depletion of ^{129}Xe relative to the modern composition[4,5], which tends to disappear in more recent samples[5,6]. To reconcile this deficit in the Archaean atmosphere by mantle degassing would require the degassing rate of Earth at the end of the Archaean to be at least one order of magnitude higher than today. We demonstrate that such an intense activity could not have occurred within a plate tectonics regime. The most likely scenario is a relatively short (about 300 million years) burst of mantle activity at the end of the Archaean (around 2.5 billion years ago). This lends credence to models advocating a magmatic origin for drastic environmental changes during the Neoarchaean era, such as the Great Oxidation Event.

The terrestrial atmosphere contains a ^{129}Xe monoisotopic excess of 7.3% relative to primordial (solar or meteoritic) xenon, attributed[7] to the decay of the extinct radioisotope ^{129}I. Some ^{129}Xe may also have been inherited from comets during the early stages of Earth's accretion[3]. Atmospheric xenon evolved subsequently through mass-dependent fractionation (MDF) due to selective atmospheric escape[4,5,8,9], while preserving the mass-independent, monoisotopic excess of ^{129}Xe. Degassing of mantle xenon through volcanism contributed further ^{129}Xe to the atmospheric inventory, because mantle xenon is enriched in ^{129}Xe. (Values of ^{129}Xe/^{130}Xe up to 7.0 are found for mantle plumes, and up to 7.8 for mid-ocean ridge basalt (MORB) mantle source[10,11], relative to the present-day atmospheric Xe signature[7], ^{129}Xe/^{130}Xe = 6.496). Remnants of ancient atmospheric gases have been identified in fluid inclusions hosted in Archaean hydrothermal quartz[5,8,9,12] and trapped in organic matter isolated from Archaean chert[4], all from Australia and South Africa. Whereas nitrogen, neon, argon (^{36}Ar, ^{38}Ar) and krypton have isotopic compositions indistinguishable from the modern atmospheric values[4,5,8,12], xenon isotopes are subject to MDF to an extent that is intermediate between the composition of the atmospheric Xe ancestor (labelled U-Xe) and the modern composition. The extent of isotope fractionation increased with time to reach the modern Xe composition[5] around 2 billion years ago (Ga). Together with the under-abundance of Xe in air relative to the expected abundance pattern of chondritic noble gases, this evolution has been attributed to selective Xe escape from the atmosphere to space via a non-thermal escape process related to interactions between the upper atmosphere's atoms and ultraviolet photons from the young Sun[5,13].

Samples with ages between 3.3 Ga and 2.7 Ga present comparable depletions of ^{129}Xe relative to adjacent ^{128}Xe and ^{130}Xe isotopes irrespective of their sampling location, whereas more recent samples have compositions consistent with that of modern atmospheric Xe (Figs. 1 and 2; Extended Data Table 1). The deficit of ^{129}Xe in the Archaean atmosphere (denoted $^{129}Xe_{DEF}$) was compensated over time by degassing of ^{129}Xe-rich mantle xenon (labelled $^{129}Xe_{XS}$, where suffix XS refers to ^{129}Xe in excess of the atmospheric composition: $^{129}Xe_{XS} = {}^{130}Xe_{mantle} \times \{(^{129}Xe/^{130}Xe)_{mantle} - (^{129}Xe/^{130}Xe)_{atm}\}$). We define Δ^{129}Xe as the deviation of the sample ^{129}Xe/^{130}Xe from the value expected for fractionated modern atmospheric xenon, in parts per thousand (‰) (Fig. 1). All available data[4–6,8] define a clear evolution from negative Δ^{129}Xe values around 3 Ga trending towards a modern-like composition starting around 2 Ga (Figs. 2 and 3). This evolution suggests that the flux of $^{129}Xe_{XS}$ from the mantle has varied considerably over time.

The amount of $^{129}Xe_{DEF}$ 3.3–2.7 Ga ago was $(2.56 \pm 1.02) \times 10^{10}$ mol (95% confidence intervals, CI), computed with an error-weighted average

[1]Centre de Recherches Pétrographiques et Géochimiques, UMR 7358 CNRS and Université de Lorraine, Vandoeuvre-les-Nancy, France. [2]Université de Paris, Institut de Physique du Globe de Paris, CNRS, Paris, France. *e-mail: bernard.marty@univ-lorraine.fr

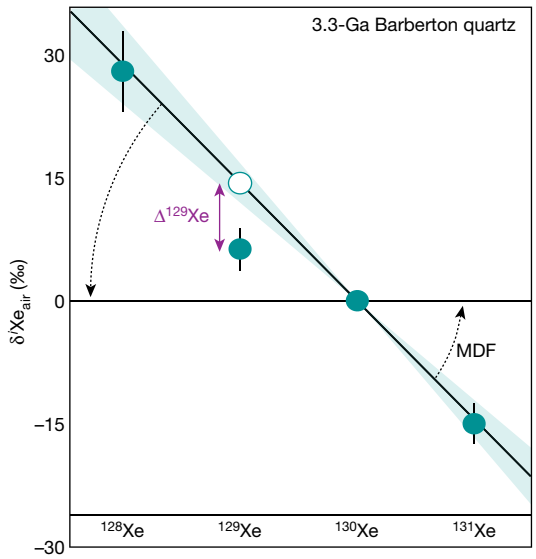


Fig. 1 | Principle of xenon isotope evolution over time. Data (filled symbols, error bars are 1σ) for fluid inclusions in 3.3-Ga Barberton (South Africa) hydrothermal quartz[8] exemplify the Xe isotope composition of Archaean air, normalized to the composition of modern air (horizontal black line). The Xe isotopic composition of the Archaean atmosphere is mass-dependently fractionated, being enriched in light isotopes relative to heavy ones. The measured ^{129}Xe/^{130}Xe ratio is, however, below the isotope fractionation line defined by the other Xe isotopes. This depletion in ^{129}Xe/^{130}Xe, denoted Δ^{129}Xe, is defined as the distance between the observed value and that expected for isotope fractionation of modern air (open symbol). The dotted arrows indicate the evolution of Xe isotope fractionation through time, yielding the modern composition around 2 Ga. The light green shading around the isotope fractionation line represents the 2σ error of the error-weighted linear correlation through Xe isotope data, excluding ^{129}Xe.

Δ^{129}Xe of (−6.3 ± 2.5)‰ (Barberton, MGTKS3#2 and Fortescue samples, Extended Data Table 1; 95% CI)). The fractionation of Xe isotopes in the ancient atmosphere strongly suggests[4,5,8,9,13] that a large fraction of xenon was lost to space between 3 Ga and 2 Ga. Considering either an exponential law[9] or a power law[4] for Xe escape results in essentially the same loss of atmospheric xenon, equivalent to 2.5 ± 0.5 times the modern atmospheric Xe inventory (Extended Data Fig. 1). By incorporating the simultaneous loss of atmospheric xenon to space during the Archaean, the total amount of $^{129}Xe_{DEF}$ could have been as high as (8.96 ± 3.57) × 10^{10} mol (Extended Data Table 2).

Delivery of cometary Xe is an unlikely process to account for the temporal evolution of atmospheric Δ^{129}Xe (Methods), and we consider volcanic degassing of $^{129}Xe_{XS}$ as the main source of Δ^{129}Xe variation. Contrary to the case of radiogenic ^{40}Ar, which was degassed from both the mantle and the continental crust through time[14], a mantle-only origin for radiogenic ^{129}Xe is certain, so that accumulation in the atmosphere directly traces time-dependent mantle degassing and convection. Since about 3 Ga, the average flux of $^{129}Xe_{XS}$ from the mantle to the atmosphere necessary to compensate for $^{129}Xe_{DEF}$ in the Archaean atmosphere is equivalent to 8.5 ± 3.4 mol yr^{-1} (closed system atmosphere[8]), or 30 ± 12 mol yr^{-1} (taking into account xenon lost to space) (95% CI). We estimate the modern flux of $^{129}Xe_{XS}$ independently by scaling the Xe/He ratio measured within mantle-derived samples to the ^{3}He mantle flux in the oceans from submarine volcanism[15], and from subaerial volcanoes[16] to be 0.9 ± 0.5 mol yr^{-1} (Methods; Extended Data Table 2). Thus modern degassing rate averaged over 3 Ga would fail by one order of magnitude to supply the amount of $^{129}Xe_{DEF}$ that was missing in the Archaean atmosphere.

We modelled the evolution of atmospheric Xe through time by MDF[4] with different functions estimating the evolution of Δ^{129}Xe (Fig. 2A–D; Methods). The model considers both cases of a closed system atmosphere and progressive escape to space, with, in the latter case, the amount of lost Xe (2.5 times the modern Xe inventory in total) being scaled to the isotopic evolution of atmospheric Xe (Methods, Extended Data Fig. 1). This model is iterative, combining progressive loss and MDF fractionation of atmospheric xenon until 2.0 Ga with the time-dependent evolution of Δ^{129}Xe (Extended Data Fig. 1). The model requires the ^{129}Xe/^{130}Xe ratio of the ancient mantle to be estimated. Because the production of substantial amounts of radiogenic ^{129}Xe would have occurred only during the first 100 million years (Myr) of Earth's history given the half-life of parent ^{129}I (15.7 Myr), the ^{129}Xe/^{130}Xe ratio could have only evolved by subduction/recycling of 'modern-like' atmospheric xenon into the mantle[8,11,17]. From mass balance, we estimate that the pre-subduction ^{129}Xe/^{130}Xe ratio of the mantle was in the range 14 ± 1 (Methods; Extended Data Fig. 2). Although the recycling history of atmospheric Xe into the mantle between 3 Ga and 1 Ga is not known, numerical modelling of Xe evolution in the mantle–atmosphere system suggests that the imprint of recycling became quantitatively important only from 1 Ga (Extended Data Fig. 3, ref. [17]). To circumvent this uncertainty, we modelled the evolution of Δ^{129}Xe in the time interval 3–1 Ga. The evolution of Δ^{129}Xe is then modelled assuming that mantle degassing decreased continuously (using exponential and power laws) since the Archaean (Fig. 2A–C). Both exponential and power laws give similar outcomes for the flux of mantle-derived ^{129}Xe to the 3-Ga atmosphere, 18 mol yr^{-1} and 63 mol yr^{-1}, respectively, without considering escape, and 64 mol yr^{-1} and 220 mol yr^{-1}, respectively, if loss to space is taken into account.

However, near-constant Δ^{129}Xe in the range 3.3–2.7 Ga followed by a stepwise change to the modern value around 2.6–2.0 Ga (Fig. 2d) strongly suggest that the Neoarchaean was punctuated by a short burst of intense magmatic activity, consistent with the evolution of mantle potential temperatures through time[18] (Fig. 3). For such a model, considering a distinct period of intense degassing between 2.6 Ga and 2.2 Ga, we calculate a peak degassing rate of 141 mol yr^{-1} for escape to space (Extended Data Table 4). For comparison, we estimate the modern flux of mantle-derived ^{129}Xe to be 6.3 ± 2.6 mol yr^{-1} (computed with a ^{130}Xe flux of 0.85 ± 0.35 mol yr^{-1}, Extended Data Table 2, and an average mantle ^{129}Xe/^{130}Xe ratio of 7.4 ± 0.4). Our estimates rely on the mantle ^{129}Xe/^{130}Xe ratio, which is taken here to be maximal at 14 (corresponding to a pre-subduction signature; Methods). If this ratio were to be lower in the ancient mantle owing to an early onset of subduction (down to potential modern values of 7–8), then our estimates of mantle Xe fluxes would be increased substantially, by up to one order of magnitude (Extended Data Fig. 6). We therefore consider our degassing rate estimates reported above to represent lower limits. Irrespective of the model chosen, it is therefore clear that substantially higher mantle fluxes are required in the Archaean. Enhanced degassing would have had a marginal effect on the atmospheric ^{40}Ar/^{36}Ar ratio (Methods; Extended Data Fig. 4) and would be difficult to detect in other Xe isotope ratios of Archaean samples (for example, the fissiogenic ones) because all Archaean samples contain an inherited or produced fissiogenic excess that is likely to mask the original atmospheric composition (Methods).

Given the incompatibility of xenon during partial melting[19], the rate of mantle degassing is related to mantle melting. However, higher concentrations of Xe in the Archaean mantle relative to its present-day budget could potentially lower the amount of mantle degassing necessary to account for the evolution of the ^{129}Xe deficit in the Archaean atmosphere. Indeed, models investigating the time evolution of the ^{3}He/^{4}He ratio of the mantle, for example, suggest mantle ^{3}He concentrations to be higher during the Archaean[20]. With our numerical simulations (Fig. 2), we estimate that 3%–39% of xenon could have been degassed from the mantle since the Archaean, irrespective of the model adopted (Extended Data Table 4). From estimates of the present-day Xe content of the mantle calculated from the ^{130}Xe/^{3}He ratio and the modern

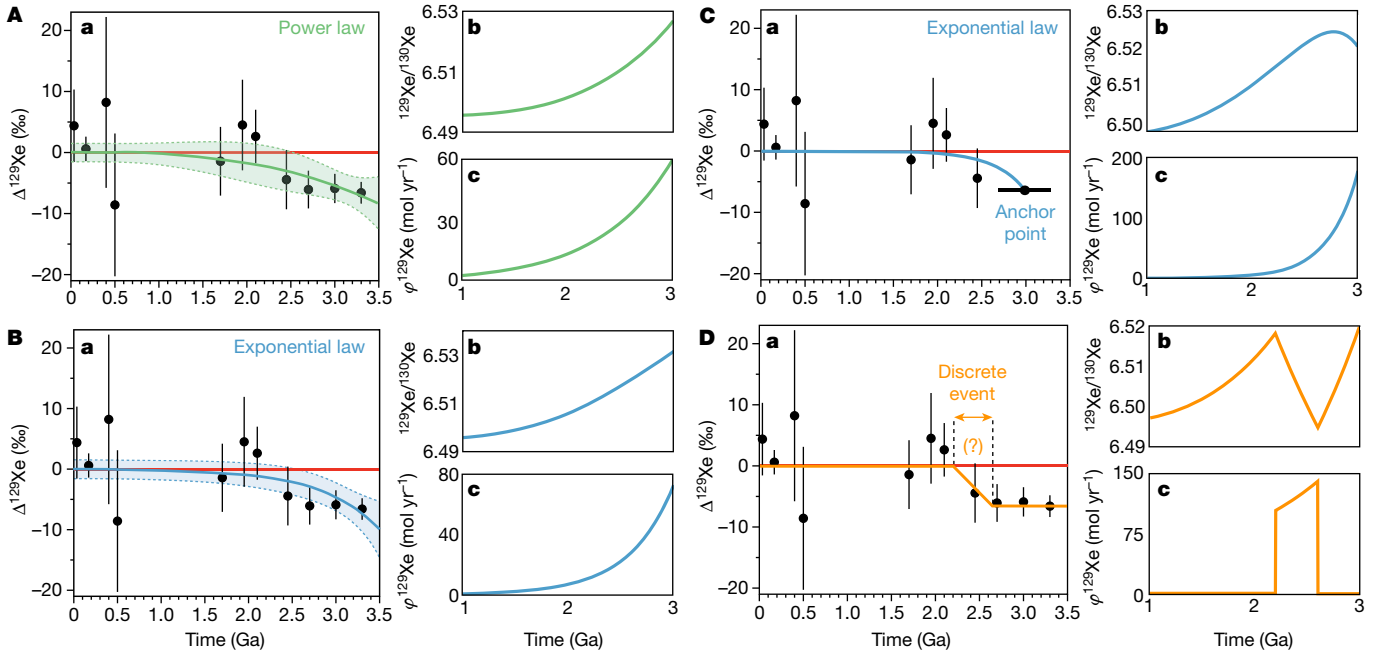


Fig. 2 | Time evolution of the deficit of ^{129}Xe (Δ^{129}Xe) in ancient atmospheric gases, of the atmospheric ^{129}Xe/^{130}Xe ratio, and of the flux of ^{129}Xe from the mantle (ϕ^{129}Xe). Data and references are given in Extended Data Table 1. The data are modelled in four ways: using power (**A**) and exponential (**B**) laws fitted through all data points; using an exponential law fitted through an anchor point at 3 Ga, −6.3‰ (**C**); and using a ramp function used to mimic the effect of a massive, discrete episode of degassing 2.6–2.2 Ga (**D**). For each of the four models **A**–**D** are shown three plots: **a**, Δ^{129}Xe versus time (data points), with a curve fitted to the data; **b**, ^{129}Xe/^{130}Xe versus time; and **c**, ϕ^{129}Xe versus time. In the plots of Δ^{129}Xe, the curves and error areas (**A**,**a**, **B**,**a**; 95% confidence interval) were produced using the error-weighted solver function of the Matlab curve-fitting tool.

^{3}He-degassing rate, we also find that the Archaean mantle Xe content could have been at best a factor of 2 higher than that of the modern mantle (Methods, Extended Data Table 2). Moderate mantle noble-gas depletion since the Archaean is independently indicated by Ne isotope systematics of the mantle–atmosphere system (Methods). We conclude that higher noble-gas concentrations in the ancient mantle cannot account for the ≥10 times greater Xe mantle fluxes in the distant past, thus calling for enhanced magma production rates in the Neoarchaean.

We evaluate here which scenario—continuous decrease in mantle degassing with time or a short-lived burst of activity in the 2.6–2.2 Ga time period—is most likely within the framework of past mantle dynamics. The former model represents the secular waning of melt generation at mid-ocean ridges and can in principle be rejected, as we now explain. In the Neoarchaean, ambient mantle temperatures were higher than today, implying larger melt fractions[18,21,22], but melt production rates depend on plate velocities, which are poorly constrained for that time. Extending plate tectonics models far back in time is fraught with severe uncertainties[18,21,22]. The global rate of plate renewal, however, is directly related to the Earth's heat loss, which can be deduced from changes of the ambient mantle temperature through time. Going back in time, this temperature increases and peaks at about 1,600 °C at an age in the 2.5–3.0 Ga range[18] (Fig. 3). Data for greater ages up to 3.5 Ga do not indicate any further temperature variation[18]. Thus, by definition, the mantle cooling rate was effectively zero at about 2.5 Ga. The global heat balance for the Earth then dictates that heat loss was equal to heat production. Given that heat production contributes about half of today's heat loss and that it was twice as large at 2.5 Ga, heat loss was about equal to its present-day value at that time. Using a well-tested model for the thermal evolution of oceanic plates, we relate the rate of melt generation to heat loss and melt thickness at spreading centres (Methods). We show that the rate of melt production cannot have been more than about five times higher than today, far below values required by the ^{129}Xe data. We thus argue in favour of a relatively short burst of mantle activity around the Archaean–Proterozoic boundary, for example, in the 2.6–2.2 Ga time period (Fig. 2d). As shown in the Methods, such intense activity would necessarily be associated with a large heat loss and would induce a dip in ambient mantle temperature if it were long-lived. The mantle ambient temperature data do not support this, and hence provide further support for a short phase of anomalous melt generation.

The large temperatures that prevailed in the late Archaean imply that large melt fractions occurred in mantle upwellings, probably affecting the global rheological behaviour of the mantle. It has been proposed that the mantle may have experienced brief 'mush ocean' episodes[21] that punctuated longer periods of sluggish plate tectonics[22]. The increased overturn rate of cooled material during 'mush oceans' would have resulted in enhanced degassing. It would also have led to larger rates of heat loss and hence rapid cooling. Thus, this anomalous convection regime was self-defeating and could not have been maintained for long. The progressive cooling of the mantle from 2.5 Ga onwards then enabled a stable plate tectonic regime[22,23] and steady-state mantle degassing (Fig. 3).

Several independent geological observations support the operation of a peculiar mantle convection regime in the late Archaean and early Proterozoic[24–26]. For example, the Superior craton saw the repeated accretion of large individual volcanic belts and older terrains at its southern margin in at least five independent events, over very short time intervals of about 10 Myr between[26] 2.70 Ga and 2.65 Ga. Following craton assembly, the very voluminous Matachewan dyke swarm testifies to enhanced magmatic activity and large eruption rates at 2.45 Ga, which are not well accounted for by plate tectonics[27]. Furthermore, the fact the ^{129}Xe/^{130}Xe of the atmosphere has not changed since 2.2 Ga, despite continued large-scale magmatism, indicates that (i) the concentration of Xe in the mantle was lowered during intense degassing periods, and/or that (ii) the initiation of subduction-driven transfer of atmospheric Xe to the mantle during the Archaean–Proterozoic diminished the difference in ^{129}Xe/^{130}Xe between the two reservoirs (Methods).

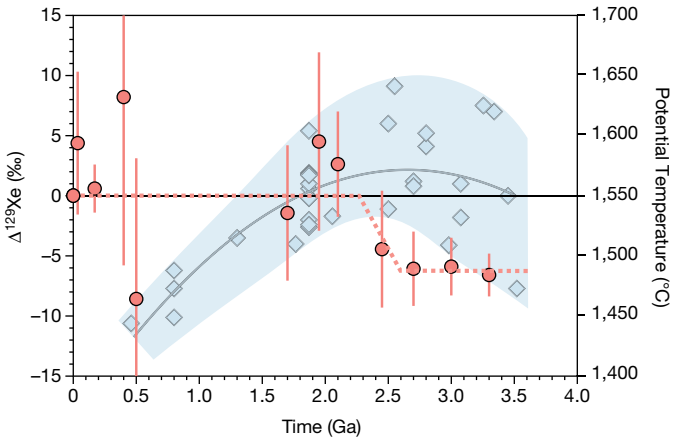


Fig. 3 | Time evolution of the deficit of ^{129}Xe (Δ^{129}Xe) in ancient atmospheric gases compared to petrological estimates of mantle potential temperature (T_P) for non-arc lavas. The data for T_P (light blue diamonds) are taken from ref. [18]. The light grey curve represents the fit through the T_P data, and the light blue shaded area exemplifies the evolution of T_P through time. The filled red circles are the Δ^{129}Xe values as defined in Fig. 1 (error bars, 1σ) and in the main text, and given in Extended Data Table 1.

The period during which Xe was most efficiently degassed from the mantle to the atmosphere (2.6–2.2 Ga) occurred at a time when the Earth was undergoing fundamental environmental changes, including the Great Oxidation Event[28]. An intense period of mantle degassing at that time may therefore have been essential in promoting the transition towards modern Earth-like conditions, which was required for the development of life[2]. Unlike Xe, other volatile elements (water, carbon and nitrogen species) only behave as incompatible elements during redox conditions akin to those of the modern mantle. The near invariance of redox-sensitive elements like vanadium indicates that the redox state of the mantle and associated basalts has remained nearly constant since the middle Archaean to the present[29]. Thus volcanic H_2O, CO_2, N_2 and SO_2 (the main volcanic gas species at low pressures[30]) would have been released into the Archaean atmosphere at rates comparable to that of $^{129}Xe_{DEF}$. The Archaean volcanic flux of CO_2, which is of the order of 6×10^{12} mol yr^{-1} at present[31], would have been in the range 10^{14}–10^{15} mol yr^{-1}, comparable to the anthropogenic flux of CO_2 (7×10^{14} mol yr^{-1}). Such high volcanic gas fluxes could have had tremendous impact on the Archaean environment, providing enormous quantities of CO_2 and SO_2 and possibly triggering the Δ^{34}S peak at that epoch[27]. Enhanced CO_2 (and associated N_2) fluxes could have played a major part in the thermal budget of the Earth's surface[32], by lowering the partial pressure of atmospheric N_2 and by triggering the production of organic matter that ultimately led to the Great Oxidation Event.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1745-7.

1. Labrosse, S. & Jaupart, C. Thermal evolution of the Earth: secular changes and fluctuations of plate characteristics. *Earth Planet. Sci. Lett.* **260**, 465–481 (2007).
2. Stüeken, E. E., Kipp, M. A., Scwieterman, E. W., Johnson, B. & Buick, R. Modeling pN_2 through geological time: implications for planetary climates and atmospheric biosignatures. *Astrobiology* **16**, 949–963 (2016).
3. Marty, B. et al. Xenon isotopes in 67P/Churyumov-Gerasimenko show that comets contributed to Earth's atmosphere. *Science* **356**, 1069–1072 (2017).
4. Bekaert, D. V. et al. Archean kerogen as a new tracer of atmospheric evolution: implications for dating the widespread nature of early life. *Sci. Adv.* **4**, eaar2091 (2018).
5. Avice, G. et al. Evolution of atmospheric xenon and other noble gases inferred from Archean to Paleoproterozoic rocks. *Geochim. Cosmochim. Acta* **232**, 82–100 (2018).
6. Meshik, A. P., Hohenberg, C. M., Pravdivtseva, O. V. & Kapusta, Y. S. Weak decay of Ba-130 and Ba-132: geochemical measurements. *Phys. Rev. C* **64**, 035205 (2001).
7. Ozima, M. & Podosek, F. A. *Noble Gas Geochemistry* (Cambridge Univ. Press, 2002).
8. Avice, G., Marty, B. & Burgess, R. The origin and degassing history of the Earth's atmosphere revealed by Archean xenon. *Nat. Commun.* **8**, 15455 (2017).
9. Pujol, M., Marty, B. & Burgess, R. Chondritic-like xenon trapped in Archean rocks: a possible signature of the ancient atmosphere. *Earth Planet. Sci. Lett.* **308**, 298–306 (2011).
10. Mukhopadhyay, S. Early differentiation and volatile accretion recorded in deep-mantle neon and xenon. *Nature* **486**, 101–104 (2012).
11. Péron, S. & Moreira, M. Onset of volatile recycling into the mantle determined by xenon anomalies. *Geochem. Persp. Lett.* **9**, 21–25 (2018).
12. Marty, B., Zimmermann, L., Pujol, M., Burgess, R. & Philippot, P. Nitrogen isotopic composition and density of the Archean atmosphere. *Science* **342**, 101–104 (2013).
13. Zahnle, K. J., Gacesa, M. & Catling, D. C. Strange messenger: a new history of hydrogen on Earth, as told by xenon. *Geochim. Cosmochim. Acta* **244**, 56–85 (2019).
14. Pujol, M., Marty, B., Burgess, R., Turner, G. & Philippot, P. Argon isotopic composition of Archaean atmosphere probes early Earth geodynamics. *Nature* **498**, 87–90 (2013).
15. Bianchi, D. et al. Low helium flux from the mantle inferred from simulations of oceanic helium isotope data. *Earth Planet. Sci. Lett.* **297**, 379–386 (2010).
16. Allard, P. Global emissions of helium-3 by subaerial volcanism. *Geophys. Res. Lett.* **19**, 1478–1481 (1992).
17. Parai, R. & Mukhopadhyay, S. Xenon isotopic constraints on the history of volatile recycling into the mantle. *Nature* **560**, 223–227 (2018): correction. 563, E28 (2018).
18. Herzberg, C., Condie, K. & Korenaga, J. Thermal history of the Earth and its petrological expression. *Earth Planet. Sci. Lett.* **292**, 79–88 (2010).
19. Heber, V. S., Brooker, R. A., Kelley, S. P. & Wood, B. J. Crystal-melt partitioning of noble gases (helium, neon, argon, krypton, and xenon) for olivine and clinopyroxene. *Geochim. Cosmochim. Acta* **71**, 1041–1061 (2007).
20. Porcelli, D. & Elliott, T. The evolution of He isotopes in the convective mantle and the preservation of high ^{3}He/^{4}He ratios. *Earth Planet. Sci. Lett.* **269**, 175–185 (2008).
21. Sleep, N. H. in *Treatise on Geophysics* **9** (ed. Stevenson, D.) 145–170 (Elsevier, 2007).
22. Korenaga, J. Thermal evolution with a hydrating mantle and the initiation of plate tectonics in the early Earth. *Earth Planet. Sci. Lett.* **116**, 1–20 (2011).
23. Tang, M., Chen, K. & Rudnick, R. L. Archean upper crust transition from mafic to felsic marks the onset of plate tectonics. *Science* **351**, 372–375 (2016).
24. Blake, T. S., Buick, R., Brown, S. J. A. & Barley, M. E. Geochronology of a Late Archaean flood basalt province in the Pilbara Craton, Australia: constraints on basin evolution, volcanic and sedimentary accumulation, and continental drift rates. *Precambr. Res.* **133**, 143–173 (2004).
25. Sleep, N. H. & Windley, B. F. Archean plate tectonics: constraints and inferences. *J. Geol.* **90**, 363–379 (1982).
26. Percival, J. A., Stern, R. A. & Skulski, T. Crustal growth through successive arc magmatism: reconnaissance U–Pb SHRIMP data from the northeastern Superior Province, *Canada. Precambr. Res.* **109**, 203–238 (2001).
27. Ciborowski, T. J. R. & Kerr, A. C. Did mantle plume magmatism help trigger the Great Oxidation Event? *Lithos* **246–247**, 128–133 (2016).
28. Holland, H. D. Volcanic gases, black smokers, and the Great Oxidation Event. *Geochim. Cosmochim. Acta* **66**, 3811–3826 (2002).
29. Canil, D. Vanadium in peridotites, mantle redox and tectonic environments: Archean to present. *Earth Planet. Sci. Lett.* **195**, 75–90 (2002).
30. Gaillard, F., Scaillet, B. & Arndt, N. T. Atmospheric oxygenation caused by a change in volcanic degassing pressure. *Nature* **478**, 229–232 (2011).
31. Marty, B. & Tolstikhin, I. N. CO_2 fluxes from mid-ocean ridges, arcs and plumes. *Chem. Geol.* **145**, 233–248 (1998).
32. Kasting, J. F. Faint young Sun redux. *Nature* **464**, 687–689 (2010).

Methods

Deficit of ^{129}Xe in Archaean air relative to modern air, reservoirs and fluxes

Δ^{129}Xe is the deviation of the sample ^{129}Xe/^{130}Xe ratio from the modern atmospheric ^{129}Xe/^{130}Xe ratio, in ‰ (Fig. 1). Modern atmospheric xenon is mass-dependently fractionated relative to ancient atmosphere[4,5,8,9]. Because atmospheric ^{129}Xe is contributed by a monoisotopic nuclear effect (the decay of ^{129}I), its variation can be identified from mass-dependent isotopic fractionation by comparison to the adjacent stable Xe isotopes. Following Pujol et al.[9], the isotope ratios are normalized to the modern Xe isotope composition, and the slopes of the fractionation trends (as well as the original data) are listed in refs. [4,5,8,9] (original Xe data are reported in https://zenodo.org/record/3378722#.Xa6cay3pNvF).

Δ^{129}Xe is the distance between the measured δ^{129}Xe value (green dot, Fig. 1) and the equivalent value sitting on the fractionation line at mass 129 (white dot, Fig. 1). Values of Δ^{129}Xe different from 0 are identified for three samples having ages around 3 Ga (Extended Data Table 1). Other samples have Δ^{129}Xe values that are not statistically different from the modern atmosphere composition. Most of these samples were analysed only once and as such the resulting errors are comparatively large. For the three samples above, we computed a mean error-weighted Δ^{129}Xe value of (−6.3 ± 2.5)‰ (95% CI) for the period 3.3–2.7 Ga.

The amount of ^{129}Xe that was missing in the Archaean atmosphere, denoted $^{129}Xe_{DEF}$, is computed from the mean Archaean Δ^{129}Xe value, the modern Xe inventory of the atmosphere, and the isotopic composition of modern atmospheric Xe (Extended Data Table 2). Taking the Archaean atmospheric ratio of 6.455 (obtained by subtracting 6.3‰ from the modern ^{129}Xe/^{130}Xe ratio, after correction for mass-dependent isotopic fractionation) instead of the modern value[7] of 6.496 would make a negligible difference compared to uncertainties in Δ^{129}Xe values. We also considered a non-conservative atmosphere from which 2.5 times the modern Xe inventory is lost to space, as suggested by the temporal evolution of Xe MDF (Extended Data Fig. 1). ^{129}Xe degassed from the mantle that changes the atmospheric Xe isotopic composition is labelled $^{129}Xe_{XS}$. The average yearly flux of $^{129}Xe_{XS}$ was simply computed by dividing $^{129}Xe_{DEF}$ by 3×10^9 yr (Extended Data Table 2).

The modern mantle flux of $^{129}Xe_{XS}$, $\varphi^{129}Xe_{XS}$, is computed as:

$$\varphi^{129}Xe_{XS} = \varphi^{3}He_{mantle} \times (^{130}Xe/^{3}He)_{mantle} \times [(^{129}Xe/^{130}Xe)_{mantle} - (^{129}Xe/^{130}Xe)_{atm}] \quad (1)$$

with $(^{129}Xe/^{130}Xe)_{atm} = 6.496$ (ref. [7]). The fluxes were computed for two different mantle sources, namely mid-ocean ridge basalt (MORB; ^{129}Xe/^{130}Xe = 7.8, ref. [11]), and mantle plume (^{129}Xe/^{130}Xe = 7.0, ref. [10]). The resulting global value, 0.89 ± 0.47 mol yr^{-1}, encompasses both estimates within uncertainties (Extended Data Table 2). Note that this value is an upper limit since it assumes a end-member ratio for mantle ^{129}Xe/^{130}Xe.

Cometary contribution

We test here the possibility that the $^{129}Xe_{DEF}$ was compensated by the delivery of cometary Xe. The analysis of volatiles released by comet 67P/Churyumov–Gerasimenko (67P/C–G) suggests[3] that comets are rich in xenon and particularly in ^{129}Xe. A large amount of ^{129}Xe could have been delivered by cometary impacts in the time interval 3.0–2.0 Ga. Assuming that the xenon data from 67P/C–G[3,33] (^{129}Xe/^{130}Xe = 7–8, Xe/$H_2O = 2.4 \times 10^{-7}$, H_2O concentration ~20 wt%, density 0.55 g cm^{-3}) are representative of the cometary reservoir, a single comet with a diameter of ~260 km impacting the Earth could have delivered the amount of ^{129}Xe missing in the Archaean atmosphere. For comparison, the impactor that made the 2.02-Ga-old Vredefort impact structure (the second largest one preserved on Earth) might have been much smaller, around 10–20 km in diameter[34]. Several cometary impacts would have resulted in a similar effect, without leaving scars on Earth if they occurred in the oceans, or if comets exploded in the upper atmosphere. However, this possibility is not consistent with the progressive isotope evolution of palaeo-atmospheric xenon, which is best accounted for by escape to space[5,8,9], whereas addition of cometary Xe would have forced Archaean atmospheric Xe towards a primitive composition rather than a modern atmospheric one. We therefore consider the addition of cometary ^{129}Xe during the Archaean to be insignificant compared to the contribution from mantle degassing.

Mantle degassing state

We evaluate here how much mantle Xe should have been lost from the 3-Ga mantle through time in order to supply missing ^{129}Xe to the atmosphere since 3 Ga ($^{129}Xe_{DEF} = 2.6 \times 10^{10}$ mol for a closed system atmosphere, and 9×10^{10} mol in the case of atmospheric escape, Extended Data Table 2). We consider two mantle sources[10,11], MORB-like (^{129}Xe/^{130}Xe = 7.8) and mantle-plume-like (^{129}Xe/^{130}Xe = 7.0). The mantle Xe contents are scaled to those of ^{3}He. The MORB source ^{3}He content is computed from the ^{3}He flux to the oceans and subaerial volcanoes[15,16], the magma generation rate at ridges (21 km^3 yr^{-1}) and an average partial melting rate of 12% (ref. [35]). The plume source content is derived from the difference in the helium isotope ratios and in the U, Th contents between MORB and plume sources (ref. [35], see ref. [36] for comparable values). Two cases are considered, the modern mantle and the ancient, pre-subduction mantle. For the latter, we use ^{129}Xe/^{130}Xe = 14 (±1) which is our estimate for pre-atmospheric contamination of mantle xenon (based on Xe isotope correlations for CO_2 well gases; compare ref. [37]; Extended Data Fig. 2), and we correct the ^{130}Xe/^{3}He ratio for 80% atmospheric contribution, assumed to have taken place quantitatively in the last billion years (see Methods section 'Numerical modelling' below). We finally compute the lost fraction for each reservoir and for each scenario (Extended Data Table 3). In all scenarios, a MORB-type reservoir would have lost between 59% and 99.4% mantle Xe. A plume source reservoir would have lost between 3.5% and 64% Xe. A pure depleted MORB-type composition at 3 Ga is unlikely given the timing of continental crust growth (which was the primary cause of mantle depletion), and a modern-like Xe isotope composition might not have prevailed before 1 Ga. Hence it may be relevant to consider a pure ocean island basalt (OIB)-like, or mixed MORB-plume composition, with atmospheric Xe recycling taking place in the last billion years, yielding a moderate 3-Ga mantle degassing state of ~50% or less. Thus the Archaean mantle could have been richer by a factor of approximately ≤2 in xenon compared to the modern mantle.

Higher concentrations of Xe in the Archaean mantle relative to its present-day budget could potentially lower the amount of mantle degassing necessary to account for the evolution of the ^{129}Xe deficit in the Archaean atmosphere. We have estimated above that the concentration of xenon could have been ≤50% higher at 3 Ga based on mass balance of the mantle reservoir. However, models investigating the time evolution of the ^{3}He/^{4}He of the mantle, for example, require mantle ^{3}He concentrations to be higher during the Archaean[20,37]. Therefore, we have attempted to define independently the maximum amount of Xe that could be present within the Archaean mantle by constraining it with Ne isotopes.

Neon provides a useful tool for constraining the concentration of Xe in different reservoirs as is not efficiently recycled to the mantle (in contrast to Ar, Kr and Xe; ref. [38]) and has been retained in the atmosphere throughout Earth's history (in contrast to He). An additional problem with scaling our calculations against He would be that, unlike the case of Ne, the isotopic composition of the mantle end-member is not known. For Ne, if the mantle was enriched in the Archaean (3.3 Ga) relative to the modern day by a factor of 10–20, as has been suggested for ^{3}He, then the progressive degassing of mantle Ne to the atmosphere with time will result in a change in the isotopic ratio of the atmosphere, as there is a discernible difference in the Ne isotopes between the solar/chondritic ^{20}Ne/^{22}Ne ratio of the mantle (12.7–13.4, refs. [39–41]) and the atmosphere (9.80). However, as of yet, no Archaean aged samples have

shown deviations in Ne isotopes from the modern atmosphere[5], indicating limited contribution of the mantle Ne signature to the atmosphere since 3.3 Ga.

We define the maximum possible Ne enrichment factor for the Archaean mantle that could still preserve the modern day atmospheric composition through time by using concentration weighted isotopic mixing calculations. First, we assume that any enrichment in Ne concentrations within the Archaean mantle relative to the present will be ultimately degassed and retained in the atmosphere. Thus, if the mantle was 10 times more enriched in Ne during the Archaean, the Archaean atmosphere must be depleted by the same amount. We take the minimum measured $^{20}Ne/^{22}Ne$ ratio measured within 3.3-Ga quartz-hosted fluid inclusions to be that of the Archaean atmosphere (9.64 ± 0.05; ref. [5]). The mantle $^{20}Ne/^{22}Ne$ is defined as either having a solar (13.4; ref. [41]) or a chondritic-like (12.7; ref. [40]) composition. We determine that to raise the $^{20}Ne/^{22}Ne$ of the Archaean atmosphere from 9.59 to the modern value of 9.8 would require 1.60×10^{14} mol of mantle ^{20}Ne to be degassed to the atmosphere assuming the mantle has solar $^{20}Ne/^{22}Ne$, and 1.96×10^{14} mol if the mantle has chondritic $^{20}Ne/^{22}Ne$. The amount of mantle neon degassed to the atmosphere since 3.3 Ga can be expressed as:

$$^{20}Ne_{degassed} = {}^{20}Ne_{MA} - {}^{20}Ne_{AA} = {}^{20}Ne_{AI}\left\{\frac{[(^{20}Ne/^{22}Ne)_{AA} - (^{20}Ne/^{22}Ne)_{MA}]}{[(^{20}Ne/^{22}Ne)_{AA} - (^{20}Ne/^{22}Ne)_{mantle}]}\right\} \quad (2)$$

where subscripts AA, MA, AI and mantle refer to Archaean atmosphere, modern atmosphere, atmospheric inventory and mantle, respectively. Readmitting the amount of degassed Ne back to the mantle would result in the mantle during the Archaean being enriched by a factor of 1.2–3.8 times the present concentrations (mantle inventories obtained from end-member mantle ^{20}Ne concentrations[36,42], and a mantle mass of 4×10^{27} g) assuming a solar mantle, and 1.2–4.4 if the mantle Ne is chondritic. The large range in these estimates is controlled primarily by the large uncertainty on the concentration of Ne in the present-day mantle[36,42].

Potential impact of Archaean degassing on atmospheric noble gases

We tested the effect of intensive mantle degassing during the Archaean on the evolution of the atmospheric $^{40}Ar/^{36}Ar$ ratio. We carried out a mass balance calculation based on the mantle noble-gas composition on one hand, and on the other, the amount of $^{129}Xe_{DEF}$ in the atmosphere. We considered two mantle sources, MORB-like and plume-like, with noble-gas end-member compositions[10,11]. The respective $^{40}Ar/^{36}Ar$ ratios were computed at 3 Ga (correcting for radiogenic ^{40}Ar produced afterwards), and we considered a pre-subduction, Archaean $^{129}Xe/^{130}Xe$ ratio of 14 (Extended Data Fig. 2). Results suggest that the contribution of Archaean mantle degassing to the ^{40}Ar atmospheric inventory was of the order of a few per cent (Extended Data Table 3). We tested the effect of 5% and 10% ^{40}Ar inventory degassing during a sudden release of ^{129}Xe at 2.6–2.2 Ga, with a K-Ar box model similar to that used by Pujol et al.[14] that includes early degassing and crustal growth. The evolution curves are depicted in Extended Data Fig. 3. In principle, a jump of the Ar isotopic ratio around that period of time could be observable, but uncertainties related to the contribution of ^{40}Ar produced in situ in samples could mask such an effect. Thus we conclude that a massive Archaean mantle degassing event would not have drastically affected the radiogenic ^{40}Ar budget of the atmosphere.

During Archaean degassing, fissiogenic Xe isotopes were also released together with $^{129}Xe_{XS}$ from the mantle to the atmosphere. In mantle-derived samples, $^{129}Xe/^{130}Xe$ correlates with $^{136}Xe/^{130}Xe$ with a slope of 3.0 for both MORB and plume sources[10,11], as a result of contributions of radiogenic ^{129}Xe and fissiogenic ^{136}Xe ($^{136}Xe_f$). An Archaean $\Delta^{129}Xe$ value of −6‰ (Extended Data Table 1) would therefore correspond to a deficit of $^{136}Xe_f$ of about −2‰ in Archaean air. Such a variation would be barely detectable in ancient samples. Archaean samples analysed so far[4,5,8] present positive $\Delta^{136}Xe_f$ values of +30‰ (Barberton sample[5]) and higher[8], with fission spectra consistent with production from ^{238}U fission[5]. Thus any potential effect of mantle degassing is likely to be masked by the inheritance of fissiogenic Xe from the trapped crustal fluids and/or the in situ production from ^{238}U fission after emplacement of the rocks at the surface. This problem would prevent detection of any effect on the Archaean atmospheric composition of fissiogenic Xe.

Numerical modelling

We consider three scenarios around the evolution curve of the isotopic composition of atmospheric Xe (Extended Data Fig. 1), which has been modelled to follow a power law defined by $y = 0.238x^{3.41}$ (ref. [4]). New data on the isotopic composition of ancient atmosphere Xe from fluid inclusions in hydrothermal quartz have been recently published[5] that support the validity of this evolution curve. Also reported in Extended Data Fig. 1 is the theoretical amount of extra ATM_{Xe} in the atmosphere scaled on Xe isotopic evolution, where ATM_{Xe} stands for the total inventory of Xe in the present-day atmosphere. Over the lifetime of the atmosphere, ~10 ATM_{Xe} would have been lost to space.

In the first scenario, we consider that Xe degassing from the mantle occurred after Xe loss to space ended. This implies that the $\Delta^{129}Xe$ remained constant (at (−6.3 ± 2.5)‰) from 3 Ga to 1 Ga, before $\Delta^{129}Xe$ was raised to 0‰ solely through mantle degassing. In this case, mantle degassing takes place while the Xe isotope signature of the atmosphere is already modern-like, with no concomitant loss to space.

In the second scenario, we consider that Xe degassing from the mantle (with $\Delta^{129}Xe$ varying from (−6.3 ± 2.5)‰ to 0‰) occurred at 3 Ga, when the atmosphere was mass dependently fractionated by −10‰ u^{-1} and had about 3.5 times the present-day inventory of atmospheric Xe (Extended Data Fig. 1, right hand *y* axis). In this case, given that Xe is more abundant in the atmosphere than in the first scenario, the total amount of mantle-derived ^{129}Xe required to fill the ^{129}Xe deficit ($^{129}Xe_{DEF}$) is also larger than in the first scenario.

The first and second scenarios do not represent real world conditions, as they assume that degassing and loss did not occur simultaneously, but they are useful in setting the boundary conditions to this model. In the third scenario, we produce an iterative model combining progressive loss and MDF of atmospheric Xe (Extended Data Fig. 1) with the evolution of $\Delta^{129}Xe$ (Fig. 2). To model the latter, we test three possibilities by fitting $\Delta^{129}Xe$ data with either power law, exponential, power-law or ramp functions (Fig. 2A–D). At each step of the iteration (*i*), the atmosphere is allowed to evolve by both loss to space and MDF (Extended Data Fig. 4). The $^{129}Xe/^{130}Xe$ ratio is then computed by using both $\Delta^{129}Xe$ (*i* − 1) and $\Delta^{129}Xe$ (*i*). The contribution of mantle-derived ^{129}Xe ($^{129}Xe_{XS}$) to the atmospheric budget of ^{129}Xe from step *i* − 1 to step *i* is then calculated given the equation:

$$M_{contrib^{129}Xe} = \frac{(^{129}Xe/^{130}Xe)_i - (^{129}Xe/^{130}Xe)_{i-1}}{(^{129}Xe/^{130}Xe)_M - (^{129}Xe/^{130}Xe)_{i-1}} \quad (3)$$

where $(^{129}Xe/^{130}Xe)_i$ and $(^{129}Xe/^{130}Xe)_{i-1}$ are the $^{129}Xe/^{130}Xe$ of the atmosphere at steps *i* and *i* − 1, respectively, and $(^{129}Xe/^{130}Xe)_M$ is the $^{129}Xe/^{130}Xe$ of the mantle source. The amount (in mol) of mantle-derived ^{129}Xe degassed into the atmosphere between step *i* − 1 and step *i* is then calculated as:

$$M_{^{129}Xe} = M_{contrib^{129}Xe} \times ATM_{^{129}Xe_i} \quad (4)$$

where $ATM_{^{129}Xe_i}$ is the total amount (in mol) of ^{129}Xe in the atmosphere at step *i* given the evolution curve of Xe loss and $^{129}Xe/^{130}Xe$ computed

from the evolution curve of atmospheric Xe isotopes and $\delta^{129}Xe_{def(i)}$. However, determining the ($^{129}Xe/^{130}Xe$) of the mantle is not straightforward, given that this ratio also evolved through time by ^{129}Xe production through radioactive decay of now extinct ^{129}I and recycling of atmospheric Xe into the solid Earth. Given that the half-life of ^{129}I is short ($T_{1/2}$ = 15.7 Myr), the whole budget of $^{129}Xe^*$ (that is, ^{129}Xe produced by the decay of ^{129}I) should have been established early in Earth's history, within the first ~100 Myr. The recycling of atmospheric Xe to the mantle is considered to be extensive, with the present-day inventory of Xe in the mantle dominated by 80%–90% recycled modern atmosphere[10,11]. Correcting the mantle $^{129}Xe/^{130}Xe$ (7.8) for the contribution of recycled atmosphere (80%–90%) would yield a $^{129}Xe/^{130}Xe$ in the range of 13–17 for the primitive convective mantle. The initial $^{129}Xe/^{130}Xe$ of the convective mantle can also be estimated from $^{128}Xe/^{130}Xe$ versus $^{129}Xe/^{130}Xe$ correlations in magmatic CO_2 well gases (ref. [43], Extended Data Fig. 2). The $^{128}Xe/^{130}Xe$ of the initial mantle is taken as the chondritic value ($^{128}Xe/^{130}Xe_{AVCC}$ = 0.5073 ± 0.0038, where suffix AVCC refers to Average Carbonaceous Chondrite, ref. [7]). Extrapolating the $^{129}Xe/^{130}Xe$ to $^{128}Xe/^{130}Xe_{AVCC}$ yields a $^{129}Xe/^{130}Xe_{initial}$ between 13 and 15, in good agreement with independent estimates from the fraction of recycled atmosphere in the mantle. Note that these estimates assume that the atmospheric Xe component in the mantle has a modern atmospheric composition. If Xe was extensively recycled to the mantle while the Xe composition of the atmosphere was still evolving, then estimating the $^{129}Xe/^{130}Xe$ of the mantle during the Archaean becomes more complicated. However, Parai and Mujhopadhyay[17] proposed that substantial full-scale recycling of atmospheric xenon into the solid Earth could not have occurred before 2.5 Ga, given that (i) the isotopic composition of atmospheric Xe progressively evolved through time by MDF and reached the modern composition around 2 Ga (Extended Data Fig. 1), and (ii) the Xe atmospheric component in the present-day mantle is indistinguishable from modern atmosphere. Although the recycling history of atmospheric Xe into the mantle between 2.5 Ga and 1 Ga is not known, constraints on the amount of Xe being transported into the solid Earth over time through atmospheric recycling have been recently set via numerical modelling of Xe evolution in the mantle–atmosphere system (Extended Data Fig. 5, ref. [17]). While some small scale recycling of atmospheric Xe to the mantle might have occurred before 2.5 Ga, it would have had a limited effect on the budget and isotopic composition of mantle Xe, and we therefore consider our estimates of mantle $^{129}Xe/^{130}Xe$ between 13 and 15 during the Archaean to be valid.

The $\Delta^{129}Xe$ evolution curves are represented in Fig. 2. We also provide the time evolution of the atmospheric $^{129}Xe/^{130}Xe$ ratio and the flux φ of mantle-derived ^{129}Xe ($\varphi^{129}Xe$) to the atmosphere. The $^{129}Xe/^{130}Xe$ ratio might not vary monotonically because two independent processes (namely MDF of the atmosphere and mantle degassing) are causing this ratio to vary (decrease and increase through time, respectively). In the case of a short burst, the $^{129}Xe/^{130}Xe$ ratio would first decrease due to MDF, and increase during the burst (during which MDF is still ongoing but mantle degassing dominates), and then decrease again because of MDF. However, the $\Delta^{129}Xe$ would either remain stable during periods of limited mantle degassing, or increase towards 0 during period(s) of intense mantle degassing. The results of the different model versions are summarized in Extended Data Table 4.

Rate of melt production by mantle convection

For a physical model of the Earth's secular thermal evolution, one needs an equation that relates heat loss to temperature. Many past efforts have been based on physical models of sea floor spreading that have been calibrated using present-day plate characteristics[1,44]. These models must be tuned to account for changes of plate structure, density and rigidity arising from the larger temperatures and amounts of melting that prevailed in the past. A major difficulty is that plate velocities and sizes vary by about one order of magnitude on modern Earth[1,22], so that extrapolating plate tectonics far back in time is uncertain. Other models of secular cooling have relied on first-principles convection calculations, but it has proven difficult to reproduce plate tectonics owing to the large lithospheric strength that must be overcome to initiate subduction. Thus, quantitative models for mantle convection in the Archaean must be regarded as tentative[22,44,45].

Here, we circumvent this difficulty by deriving a general relationship between heat flux and melt production rate. We assume that the degassing of basaltic melts proceeds to completion so that the degassing rate is proportional to the melt generation rate, which leads to a lower bound on the melting rate. A general equation for the surface heat flux can be written for all convection regimes save the 'stagnant lid' one. In the latter regime, convection develops below a rigid layer that caps the whole planet and does not allow surface motions. For all other regimes, the basic principle is that cooling is effected in a thermal boundary layer at the top of the mantle, where vertical velocities are negligible. Heat loss is therefore due to conduction and depends on the residence time of material at the surface. This principle has been thoroughly tested in laboratory experiments and numerical calculations, as well as on the current oceanic plates[45]. Denoting the mantle potential temperature by T_P, surface temperature by T_S, the total heat loss due to convection is:

$$Q_0 = \psi(f) S_o k \frac{T_P - T_S}{\sqrt{\pi \kappa \tau_M}} \quad (5)$$

where k is thermal conductivity, κ is thermal diffusivity and S_o is the total surface of oceanic plates involved in convective motions. A key parameter is τ_M, which is the maximum age of sea floor at the Earth's surface. $\psi(f)$ depends on the distribution of sea floor ages, which is described by some function f:

$$f(t/\tau_M) = \mathrm{d}S/\mathrm{d}t \quad (6)$$

which is the surface increment between ages t and $t + \mathrm{d}t$. The current age distribution on Earth is 'triangular', such that it decreases linearly from a maximum at $t = 0$ to zero at $t = \tau_M$. This contrasts with standard convective systems for which the age distribution is 'rectangular'[40], that is, constant between $t = 0$ and $t = \tau_M$. The difference between the two distributions has a small impact on factor $\psi(f)$ in equation (4), which is not important for this discussion. The rate of sea floor generation, denoted C_A, is such that:

$$S_o = C_A \tau_M \int_0^1 f(u)\mathrm{d}u \quad (7)$$

The age distribution has again a minor impact on the result (a maximum factor of two). The thickness of melt produced, denoted H, may be calculated from thermodynamics[45] as a function of the mantle potential temperature T_P for a given mantle composition. The volume of melt produced per unit time in a plate tectonic regime is equal to:

$$\varphi = C_A H \quad (8)$$

This leads to a relationship between the melt production rate and heat loss:

$$\varphi = \lambda \frac{Q_o H}{(T_P - T_S)\sqrt{\kappa \tau_M}} \quad (9)$$

where λ is a constant.

We may now evaluate the conditions that are needed for a ≥10-fold change in melting rate. As shown in the main text, the Archaean heat flux was about equal to today's value. The mantle temperature was about 200 °C higher than today but this only implies an ~20% change of the overall temperature contrast ($T_P - T_S$), which does not change the present argument. According to ref. [21], the thickness of melt produced in hot Archaean mantle was in the range 25–35 km, corresponding to at least a threefold increase with respect to the present-day value. In order to achieve a ≥10-fold increase in melt production rate, the maximum age of oceanic plates would need to be decreased by a factor of at least $3^2 = 9$. Today, this maximum age is 180 Myr and it is not clear how plate tectonics could have operated over less than 20 Myr in the Archaean.

We have focused on the heat flow through oceanic plates and have not discussed the potential influence of continents. This is not needed here for the following reason. Continental heat flow is very close to the amount of heat released by radioactive decay in crustal rocks[45], so that our conclusion that the Archaean oceanic heat flow had to be about equal to mantle heat production still stands.

A short-lived burst of mantle activity

Enhanced degassing necessarily implies enhanced melting and heat loss, and hence enhanced cooling of the mantle, which must lead to a dip of mantle temperature if it is maintained for a long time. The thermal impact of a pulse of high mantle activity may be difficult to detect, however. Mantle temperatures have been determined with a precision of about ±60 °C (ref. [46]) at time steps of a few hundred million years and exhibit scatter (about 100–150 °C) for ages older than 2.0 Ga (ref. [21]). With the current net energy loss of Earth (equal to heat loss minus heat production), it takes about one billion years for the mantle temperature to drop by 100 °C. A net energy loss that is ten times larger would lead to the same temperature drop in 100 Myr, at the detection limit of the current temperature data. Combining heat balance arguments with constraints from ambient mantle temperatures and ^{129}Xe data should allow tight bounds to be set on the intensity and duration of anomalous mantle activity (Fig. 3).

Data availability

The sample description and Xe data are available at https://zenodo.org/record/3378722#.Xa6cMi3pNvE.

Code availability

The Matlab code for modelling the degassing rate of Xe from the mantle is available at https://zenodo.org/record/3381874#.Xa6cey3pNvE.

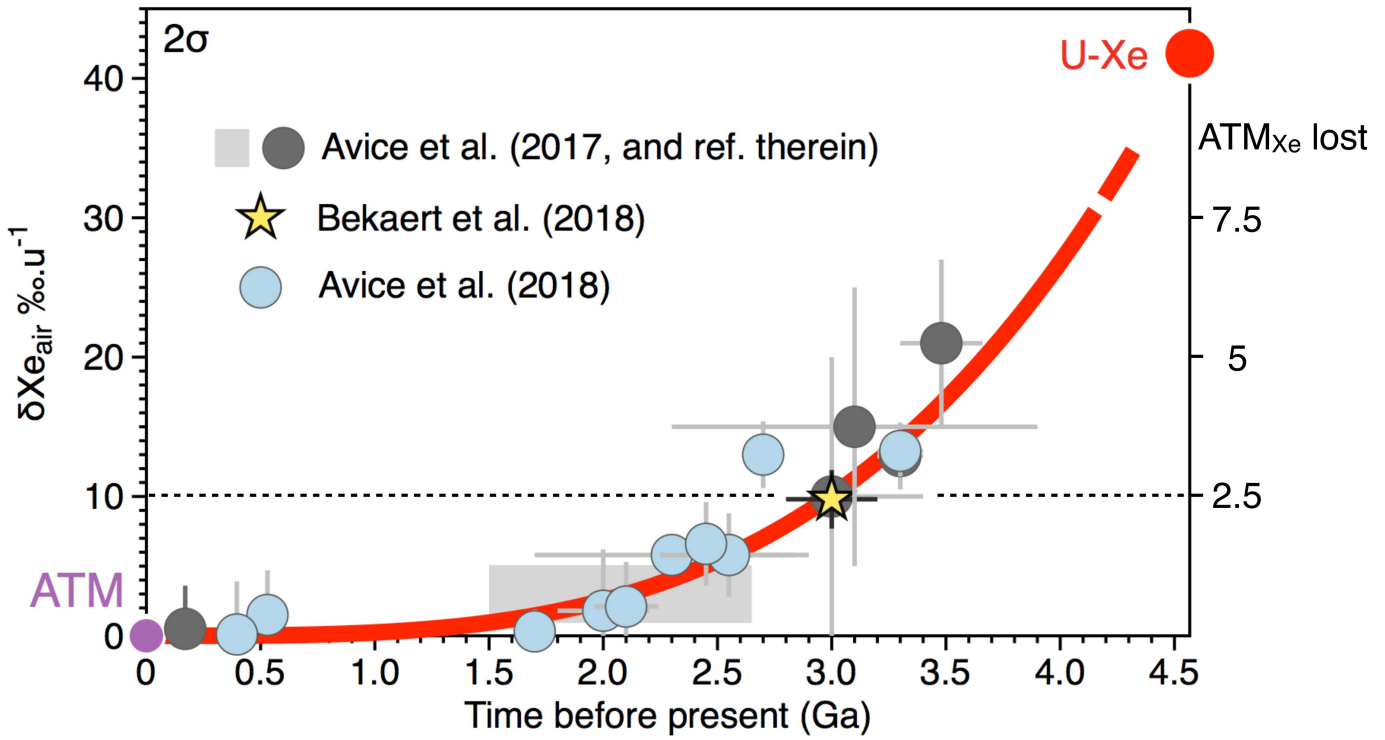
33. Rubin, M. et al. Krypton isotopes and noble gas abundances in the coma of comet 67P/Churyumov-Gerasimenko. *Sci. Adv.* **4**, eaar6297 (2018).
34. Reimold, W. U. & Gibson, R. L. Geology and evolution of the Vredefort impact structure, South Africa. *J. Afr. Earth Sci.* **23**, 125–162 (1996).
35. Tolstikhin, I. N. & Marty, B. The evolution of terrestrial volatiles: a view from helium, neon, argon and nitrogen isotope modelling. *Chem. Geol.* **147**, 27–52 (1998).
36. Halliday, A. N. The origins of volatiles in the terrestrial planets. *Geochim. Cosmochim. Acta* **105**, 146–171 (2013).
37. Gonnermann, H. M. & Mukhopadhyay, S. Preserving noble gases in a convective mantle. *Nature* **459**, 560–563 (2009).
38. Holland, G. & Ballentine, C. J. Seawater subduction controls the heavy noble gas composition of the mantle. *Nature* **441**, 186–191 (2006).
39. Ballentine, C. J., Marty, B., Sherwood Lollar, B. & Cassidy, M. Neon isotopes constrain convection and volatile origin in the Earth's mantle. *Nature* **433**, 33–38 (2005).
40. Moreira, M. & Charnoz, S. The origin of the neon isotopes in chondrites and on the Earth. *Earth Planet. Sci. Lett.* **433**, 249–256 (2016).
41. Williams, C. D. & Mukhopadhyay, S. Capture of nebular gases during Earth's accretion is preserved in deep-mantle neon. *Nature* **565**, 78–81 (2019).
42. Marty, B. The origins and concentrations of water, carbon, nitrogen and noble gases on Earth. *Earth Planet. Sci. Lett.* **313–314**, 56–66 (2012).
43. Caffee, M. W. et al. Primordial noble gases from Earth's mantle: identification of a primitive volatile component. *Science* **285**, 2115–2118 (1999).
44. Korenaga, J. in *Archean Geodynamics and Environments* (eds Benn, K., Mareschal, J. C. & Condie, K. C.) 7–32 (Geophys. Monogr. Ser. **164**, AGU, 2006).
45. Jaupart, C. & Mareschal, J.-C. in *Treatise on Geophysics* **6** (ed. Stevenson, D.) 217–251 (Elsevier, 2007).
46. Herzberg, C. & Asimow, P. D. PRIMELT3 MEGA.XLSM software for primary magma calculation: peridotite primary magma MgO contents from the liquidus to the solidus. *Geochem. Geophys. Geosyst.* **16**, 563–578 (2015).

Acknowledgements This study was supported by the European Research Council (Photonis Advanced Grant no. 695618). We thank G. Avice for discussions, and the Fondation des Treilles for providing a congenial environment in which to discuss these concepts with colleagues.

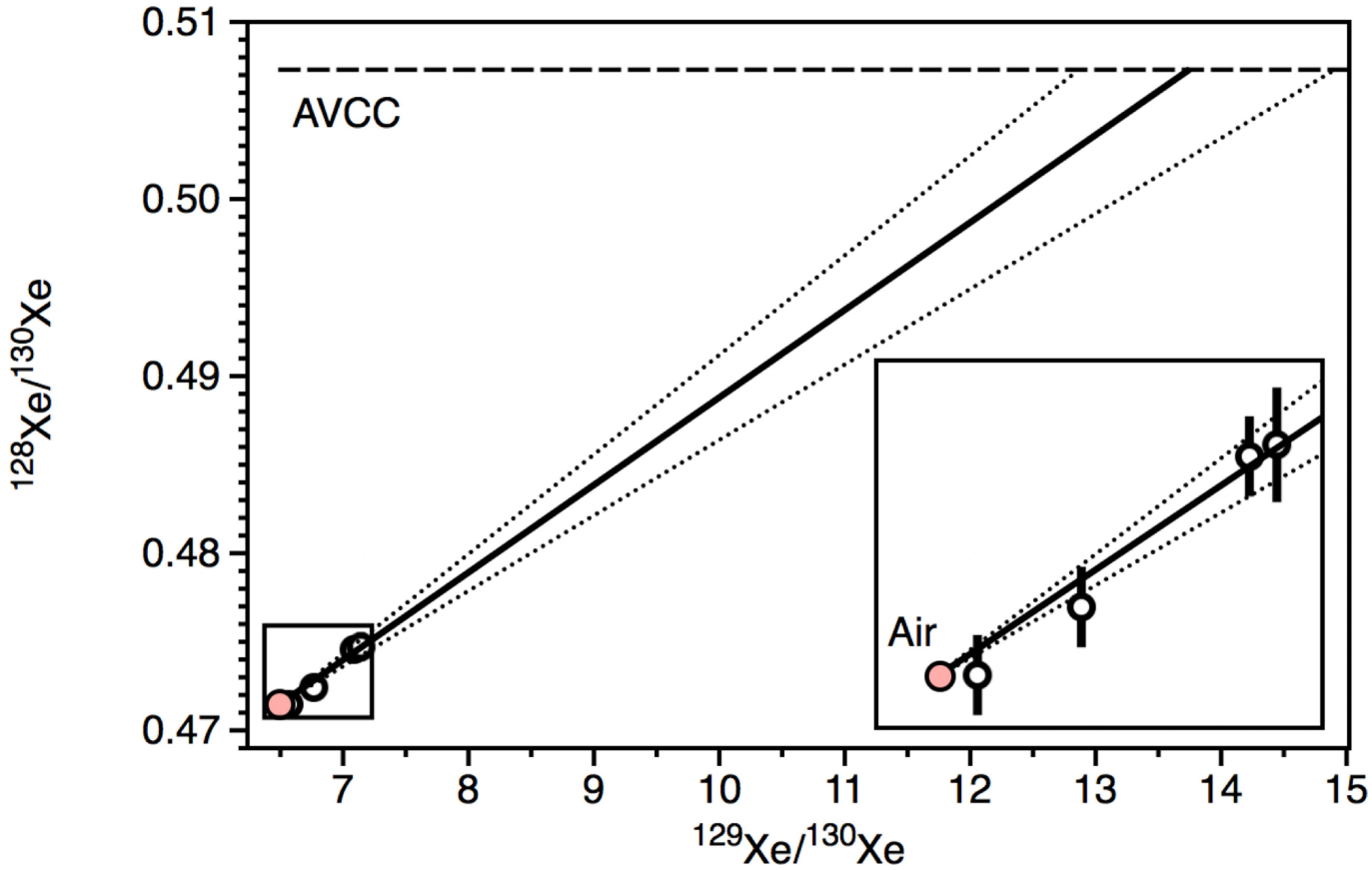
Author contributions B.M. developed the ideas presented in the manuscript. D.B. wrote the Matlab script for the numerical modelling, M.B. contributed the geological context and developed noble-gas constraints on atmospheric evolution. C.J. developed the thermal model. All authors participated in the writing of the manuscript.

Competing interests The authors declare no competing interests.

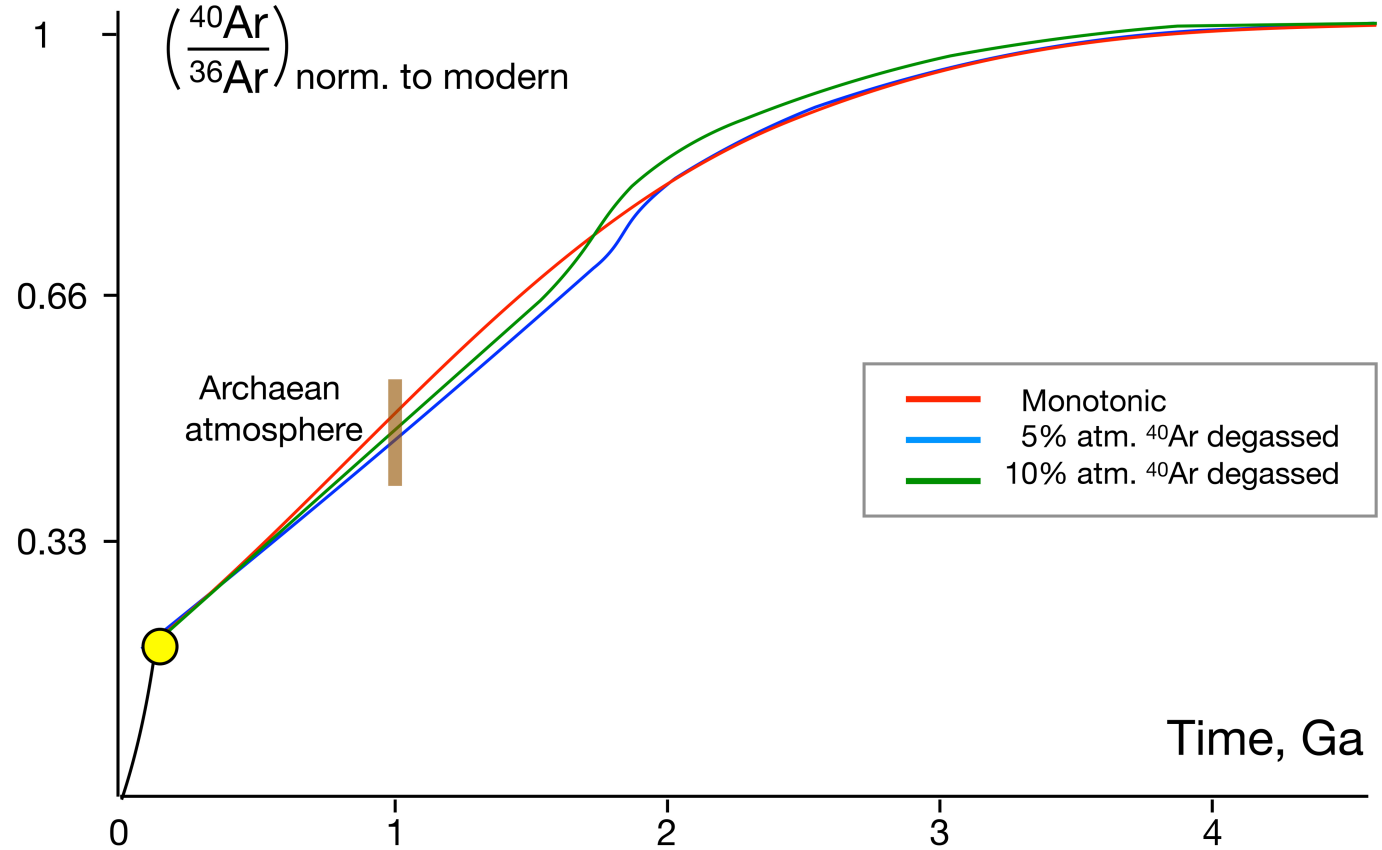
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1745-7.
Correspondence and requests for materials should be addressed to B.M.
Peer review information *Nature* thanks Greg Holland and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



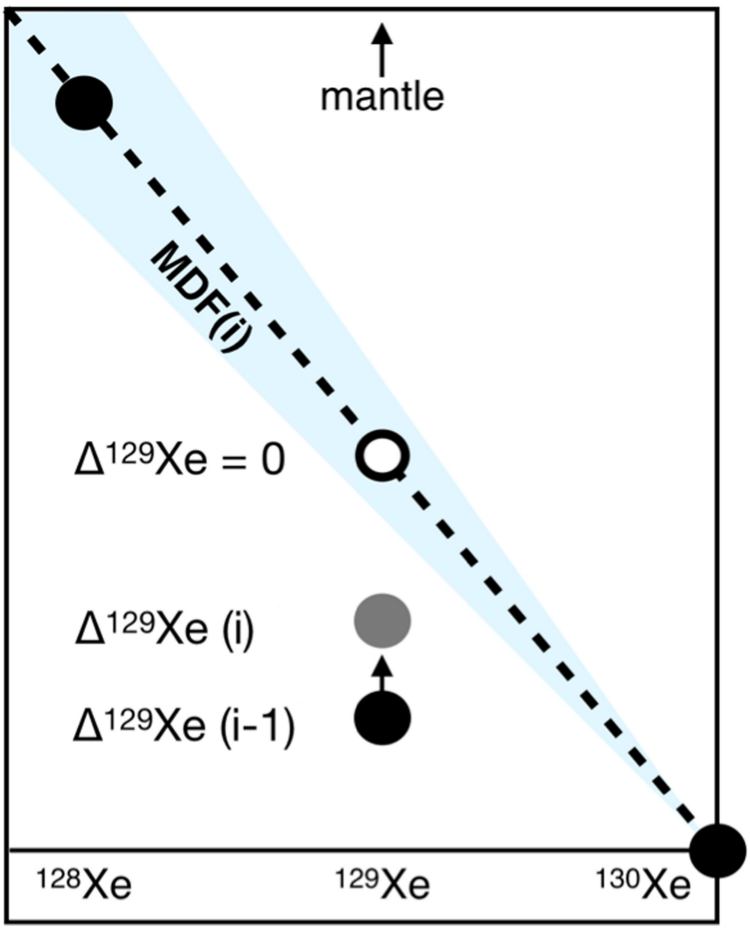
Extended Data Fig. 1 | MDF of atmospheric Xe with time relative to the modern atmosphere. Grey and blue data points[5,8] define the evolution (red curve) of atmospheric Xe mass-dependent fractionation (MDF)[4]. The left-hand *y* axis shows the isotopic fractionation of atmospheric Xe (δXe_{air}) in units of ‰ per atomic mass unit (u). The right-hand *y* axis represents multiples of the Xe inventory of the modern atmosphere, ATM_{Xe}. Error bars, ±2σ. The purple point on the left-hand side (ATM) is the modern atmospheric composition, the red dot on the right-hand side (U-Xe) is the primordial composition of atmospheric xenon[3], the grey-shaded area shows the data range from ref. [5] and references therein, and the dotted horizontal line gives the MDF value; the ATM_{Xe} values correspond to a mean age of 3 Ga.



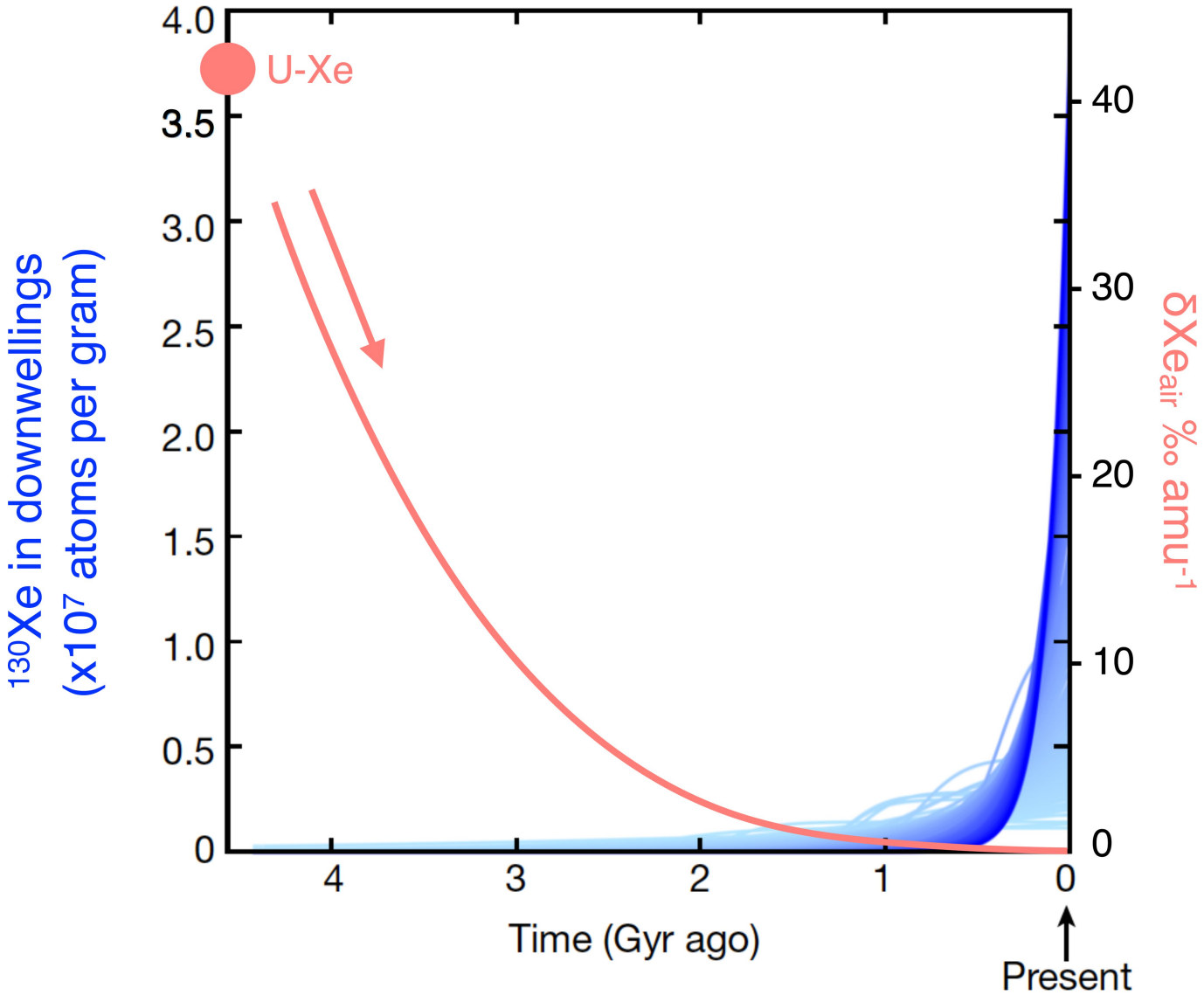
Extended Data Fig. 2 | Plot of $^{128}Xe/^{130}Xe$ versus $^{129}Xe/^{130}Xe$ for CO_2 well gases. Open circle data points are from ref. [38], and the pink filled circle shows the isotopic composition of air (error bars, 1σ). The boxed area at lower left is shown magnified in the inset. There is a correlation between the excess ^{129}Xe and ^{128}Xe (thick line, dotted thin lines define the error envelope, 95% CI) that can be used to extrapolate the primordial $^{129}Xe/^{130}Xe$ of the mantle source for an AVCC-like $^{128}Xe/^{130}Xe$ (dashed black line).



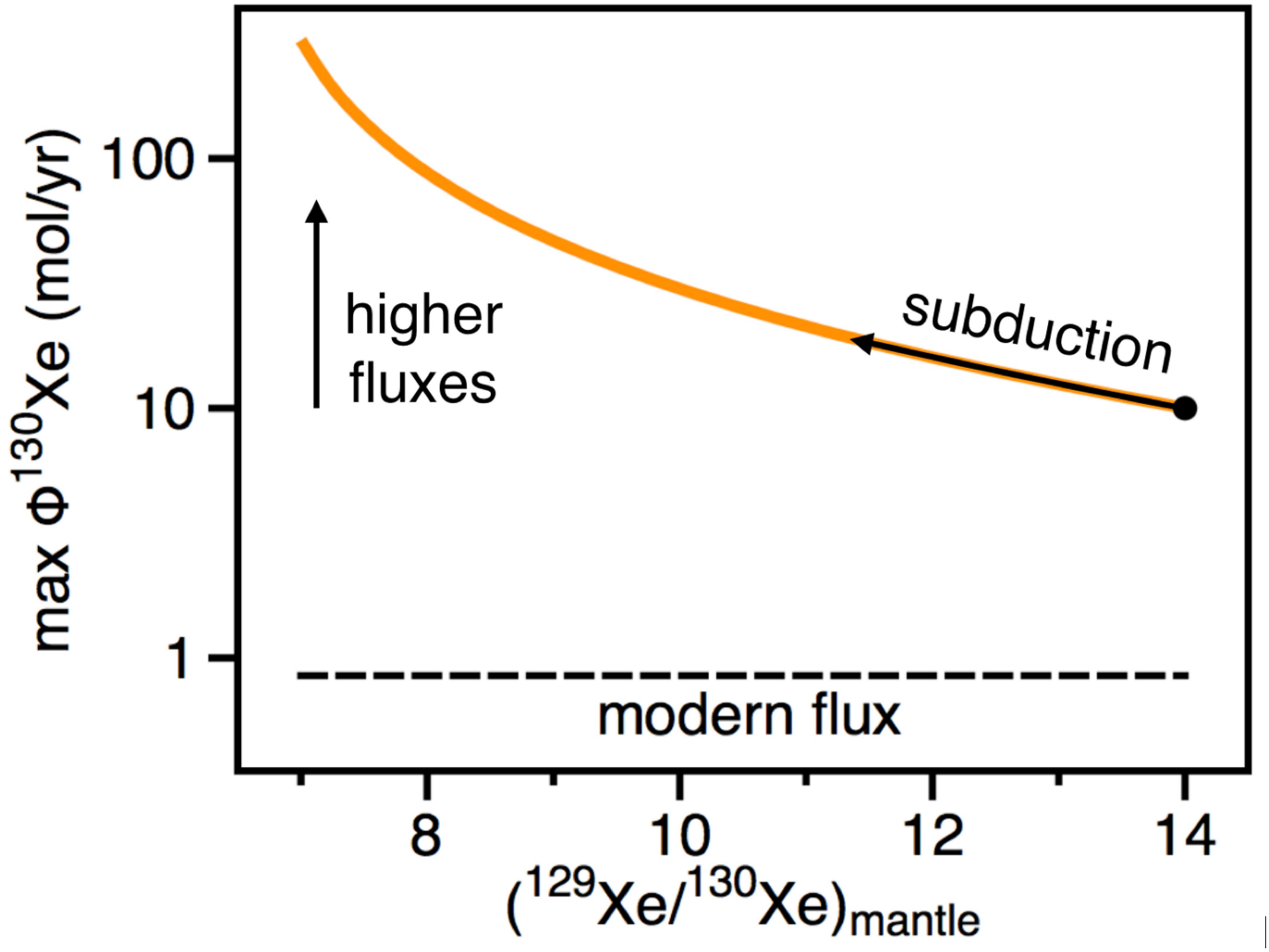
Extended Data Fig. 3 | Modelled evolution of the atmospheric $^{40}Ar/^{36}Ar$ ratio as a function of time following a mantle degassing event between 2.6 Ga and 2.2 Ga. The values are scaled to $^{129}Xe_{DEF}$, and are shown with different contributions of mantle ^{40}Ar: 0% ('Monotonic'), 5% and 10%. Atmospheric $^{40}Ar/^{36}Ar$ ratios are normalized to the present-day value of 298.6, and the evolution curves were adjusted in order to yield the modern value. The Archaean atmosphere's value is from ref. [14]. The yellow dot marks the end of catastrophic degassing and the start of continuous degassing, following ref. [14].



Extended Data Fig. 4 | Schematic representation of the method used to calculate the contribution of mantle-derived ^{129}Xe to the atmospheric budget of ^{129}Xe from step *i* − 1 to step *i*. The format is the same as in Fig. 1, where the *y* axis corresponds to δXe_{air} (only indicative here).The 'mantle' arrow indicates that the $^{129}Xe/^{130}Xe$ of the mantle end-member is high, and would plot off-graph in this space. $\Delta^{129}Xe$ values at each step of the simulation are reported on the left of the corresponding data points. The dashed line corresponds to the MDF line, with the shaded blue area representing the corresponding error envelope.



Extended Data Fig. 5 | Time series showing possible scenarios of mantle regassing histories. Shown is recycling of atmospheric Xe into the mantle (blue lines, left-hand *y* axis)[17] compared to the time evolution of atmospheric Xe isotopic composition[4] (δXe_{air}, right-hand *y* axis). The pink arrow shows the direction of atmospheric Xe isotopic evolution, from U-Xe (the progenitor of atmospheric Xe) to present. This illustrates the fact that regassing of atmospheric Xe into the mantle would have become efficient only after atmospheric Xe had reached a modern-like isotopic composition, that is, within the last 1.5 Gyr. Adapted from ref. [17].



Extended Data Fig. 6 | Maximum flux of Xe (represented by ^{130}Xe) degassed from the mantle as a function of the mantle ^{129}Xe/^{130}Xe ratio. Computations reported in Fig. 2 of the main text have been carried out using a fixed mantle ^{129}Xe/^{130}Xe of 14. Here we show that lowering this ratio (for example, via subduction of atmospheric Xe) down to modern mantle-like ^{129}Xe/^{130}Xe ≈ 7–8 would result in even greater Xe fluxes from the Archaean mantle (see black arrow). Given that the onset of atmospheric Xe recycling into the mantle is not well known, possible ^{130}Xe flux values from the Archaean mantle are within the range 10–150 mol yr^{-1} (orange curve), well above the modern flux value (0.85 ± 0.35 mol yr^{-1}, horizontal dashed line; Extended Data Table 2).

Extended Data Table 1 | $\Delta^{129}Xe$ values (in ‰) versus ages (in Ga)

	Age	$\Delta^{129}Xe$	1σ	Ref.
Barberton	3.3	-6.59	1.78	8
MGTKS3#2	3.0	-5.89	2.4	4
Fortescue	2.7	-6.07	3.09	5
Vetreny	2.45	-4.44	4.86	5
Gaoua	2.1	2.63	4.38	5
Carnaiba	1.95	4.51	7.42	5
Caramal	1.7	-1.43	5.63	5
Avranches	0.5	-8.58	11.7	5
Rhynie	0.4	8.21	14	5
Recent	0.035-0	4.38	5.94	5
Barite	0.17	0.6	2	6
air	0	0		

Sample names (left column), locations, ages and original Xe data can be found in refs. [4–6,8] and at https://zenodo.org/record/3378722#.Xa6cMi3pNvE.

Extended Data Table 2 | Archaean atmospheric inventory and modern mantle flux

Atmospheric ^{129}Xe inventory			
	Value	**±**	**Refs./notes**
Atm. Xe	1.54×10^{13}	1.77×10^{11}	7
Modern atm. $^{129}Xe/^{130}Xe$	6.496	-	7
$^{129}Xe/^{130}Xe$ with 6.3±2.5 ‰ deficit	6.455	0.016	this work
$^{129}Xe_{DEF}$ deficit in the atmosphere	2.56×10^{10}	1.02×10^{10}	95% CI
Average $^{129}Xe_{XS}$ flux over 3.0 Ga	**8.5**	**3.4**	**mole/yr**
Atm. Xe lost to space			
3.5 times modern Xe inventory	8.96×10^{10}	3.57×10^{10}	95% CI
Average $^{129}Xe_{XS}$ flux over 3.0 Ga	**30**	**12**	**mole/yr**
Modern $^{129}Xe_{XS}$ flux from the mantle			
	Value	**±**	**Refs./notes**
^{3}He flux from mantle to oceans	527	102	15
^{3}He flux from subaerial volcanism	275	**35**	16
Global ^{3}He flux from the mantle	802	137	**mole/yr**
Mantle $^{3}He/^{130}Xe$	950	**50**	10
^{130}Xe flux	**0.85**	**0.32**	**mole/yr**
MORB mantle $^{129}Xe/^{130}Xe$	7.8		10,11
MORB mantle $^{129}Xe_{XS}$ flux	1.11	0.42	
Plume mantle $^{129}Xe/^{130}Xe$	7		10,11
Plume mantle $^{129}Xe_{XS}$ flux	0.44	0.17	
Modern $^{129}Xe_{XS}$ flux	**0.89**	**0.47**	**mole/yr**

Top, atmospheric inventory of missing ^{129}Xe in the Archaean atmosphere ($^{129}Xe_{DEF}$). Bottom, modern mantle $^{129}Xe_{XS}$ flux. Xe isotope fractionation in modern air indicates[5,8] specific loss of Xe from the atmosphere to space from 4.5 Ga to about 2.0 Ga. The amount of ^{129}Xe lost to space between 3.0 Ga and about 2.0 Ga is equal to 2.5 times the modern inventory (Extended Data Fig. 1). For the modern mantle $^{129}Xe_{XS}$ flux, we considered a mantle made of 1/3 plume source and 2/3 MORB source, which yields a value intermediate between those computed for either a plume composition or a MORB composition, respectively. Data are from refs.[7,10,11,15,16].

Extended Data Table 3 | Mantle and atmosphere inventories

Mantle			
	MORB source	**Plume source**	**Notes/Refs**
Modern mantle			
^{3}He mantle source	1.50×10^{-15}	1.50×10^{-13}	35
^{130}Xe/^{3}He	1.05×10^{-3}	1.05×10^{-3}	10
Mantle ^{130}Xe (mass: 4.10^{27} g)	6.32×10^{9}	6.32×10^{11}	
^{129}Xe/^{130}Xe	7.8	7.0	
Mantle Xe	1.90×10^{9}	1.76×10^{11}	
% degassed: Closed atm	97.8%	34.2%	
% degassed: Atm. escape	99.4%	64.5%	
Archean mantle			
^{3}He mantle source	1.50×10^{-15}	1.50×10^{-13}	35
^{130}Xe/^{3}He	2.11×10^{-4}	2.11×10^{-4}	10
Mantle ^{130}Xe (mass: 4.10^{27} g)	1.26×10^{9}	1.26×10^{11}	
^{129}Xe/^{130}Xe	14	14	This work
Mantle Xe	4.05×10^{10}	5.66×10^{12}	
% degassed: Closed atm	**59.1%**	**3.5%**	
% degassed: Atm escape	**67.7%**	**5.4%**	
Atmosphere			
	MORB source	**Plume source**	**Notes/Refs**
^{130}Xe/^{3}He	1.05×10^{-3}	1.05×10^{-3}	10
^{3}He/^{36}Ar	0.5	0.8	10
^{130}Xe/^{36}Ar	8.16×10^{-4}	8.16×10^{-4}	
^{40}Ar/^{36}Ar	40,000	10,000	10,11
^{40}Ar/^{36}Ar at 3 Ga	7584	1896	
^{130}Xe/^{40}Ar	1.08×10^{-7}	4.31×10^{-7}	
^{129}Xe/^{130}Xe	14	14	Pre-subduction
$^{129}Xe_{XS}/^{40}$Ar	8.07×10^{-7}	3.23×10^{-6}	
$^{129}Xe_{DEF}$	9.0×10^{10}	9.0×10^{10}	Atm. escape
^{40}Ar degassed	1.11×10^{17}	2.79×10^{16}	
^{40}Ar atm	1.65×10^{18}	1.65×10^{18}	7
% ^{40}Ar atm	**6.8%**	**1.7%**	Contrib. to atmosphere

Concentrations are given in mol g^{-1} and abundances are given in mol. Modern ^{130}Xe/^{3}He ratios for MORB and plume sources are similar within[10] 10%. Modern mantle fluxes and concentrations are from ref. [35] and noble-gas compositions are from refs. [7,10,11]. Xe contents are computed from ^{130}Xe percentage for the different Xe isotope compositions. The pre-subduction mantle composition uses extrapolated ^{129}Xe/^{130}Xe ratios (Extended Data Fig. 2 and Methods) and, for Xe abundance, is obtained by removing 80% atmospheric Xe, corresponding to atmospheric contamination that took place during the Proterozoic. The ^{40}Ar/^{36}Ar ratios at 3 Ga are corrected for ^{40}K production during the last 3 Ga.

Extended Data Table 4 | Results of models for flux evolution through time

	$^{129}Xe_{DEG}$ (moles)	φ_{max} (mole/yr)	$\%^{130}Xe_{DEG}$	
			Marty 2012	Halliday 2013
Scenario 1 (Lower limit)	2.21×10^{10}	-	1.5%	19%
Scenario 2 (Upper limit)	7.87×10^{10}	-	5.1%	45%
Scenario 3, no escape				
Power law	1.81×10^{10}	18 (3 Ga)	1.2%	16%
Exponential 1	1.56×10^{10}	25 (3 Ga)	1.1%	14%
Exponential 2	2.23×10^{10}	63 (3 Ga)	1.5%	19%
Ramp (2.6-2.2) Ga	2.23×10^{10}	56 (2.6 Ga)	1.5%	19%
Ramp (2.5-2.4) Ga	2.23×10^{10}	224 (2.5 Ga)	1.5%	19%
Scenario 3 with escape				
Power law	3.98×10^{10}	64 (3 Ga)	2.7%	29%
Exponential 1	3.77×10^{10}	89 (3 Ga)	2.5%	28%
Exponential 2	6.10×10^{10}	220 (3 Ga)	4.0%	39%
Ramp (2.6-2.2) Ga	4.85×10^{10}	141 (2.6 Ga)	3.2%	34%
Ramp (2.5-2.4) Ga	5.02×10^{10}	519 (2.5 Ga)	3.3%	34%

See Methods for definition of scenarios. $^{129}Xe_{DEG}$ corresponds to integrated amounts (in mol) of ^{129}Xe required to have been degassed from the mantle to account for the isotopic evolution of atmospheric Xe. Φ_{max} are the maximal fluxes of mantle ^{129}Xe. $^{130}Xe_{DEG}$ (%) corresponds to the fraction of the mantle ^{130}Xe inventory required to have been degassed, for mantle budget estimates by Marty[42] and Halliday[36], and for a mantle $^{129}Xe/^{130}Xe$ of 14.

Article

A new Miocene ape and locomotion in the ancestor of great apes and humans

https://doi.org/10.1038/s41586-019-1731-0

Received: 10 July 2019

Accepted: 27 September 2019

Published online: 6 November 2019

Madelaine Böhme[1,2*], Nikolai Spassov[3], Jochen Fuss[1,2], Adrian Tröscher[2], Andrew S. Deane[4], Jérôme Prieto[5], Uwe Kirscher[1,6], Thomas Lechner[1,2] & David R. Begun[7]

Many ideas have been proposed to explain the origin of bipedalism in hominins and suspension in great apes (hominids); however, fossil evidence has been lacking. It has been suggested that bipedalism in hominins evolved from an ancestor that was a palmigrade quadruped (which would have moved similarly to living monkeys), or from a more suspensory quadruped (most similar to extant chimpanzees)[1]. Here we describe the fossil ape *Danuvius guggenmosi* (from the Allgäu region of Bavaria) for which complete limb bones are preserved, which provides evidence of a newly identified form of positional behaviour—extended limb clambering. The 11.62-million-year-old *Danuvius* is a great ape that is dentally most similar to *Dryopithecus* and other European late Miocene apes. With a broad thorax, long lumbar spine and extended hips and knees, as in bipeds, and elongated and fully extended forelimbs, as in all apes (hominoids), *Danuvius* combines the adaptations of bipeds and suspensory apes, and provides a model for the common ancestor of great apes and humans.

Many studies since the nineteenth century have investigated the origin of human bipedalism. From Darwin and Huxley to the present, many researchers have added insights into this question but with little or no fossil evidence in support[2–4]. Although many fossils have been discovered, none has shed light directly on this central question in palaeoanthropology.

Since the 1970s, many fossil apes from the middle to late Miocene epoch (13–5.3 million years ago (Ma)) from Europe have been discovered and described, along with smaller samples from the same time period in Africa[5–7]. Apes and humans are thought to have diverged at this time[8]. Some of these discoveries include partial skeletons[9,10], but none shows preservation of completely intact long bones. Although opinions vary as to the relationship of these hominids to living hominids, nearly all researchers recognize European late Miocene apes as hominids as opposed to the stem hominoids of the early and middle Miocene epoch of Africa[6,11,12].

Postcranially, the most complete fossils from Europe include the well-preserved remains of the small bones of the hand, fragments of the long bones of the limbs, a partial pelvis and partially preserved vertebrae. These discoveries have provided insights into the anatomy of late Miocene apes. We know that these apes, including *Pierolapithecus*, *Dryopithecus*, *Hispanopithecus* and *Rudapithecus*, were suspensory and similar to modern great apes to varying degrees. However, without complete long bones of the limbs and well-preserved joint surfaces (especially of the lower limbs), interpretations of details of the positional behaviour of these apes remain limited.

Reconstructing the ancestral form of positional behaviour of great apes and humans is best accomplished through the analysis of fossils. On the basis of comparisons of *Ardipithecus*, extant catarrhines and Miocene apes, it has been argued that human bipedalism evolved from a form of arboreal quadrupedalism in the last common ancestor of great apes and humans[13,14]. Others have argued that bipedalism arose from a more suspensory ancestor, based largely on fossil evidence of late Miocene hominids[6,11]. These scenarios are based on fragmentary fossil evidence. Here we present a different scenario based on our analysis of a well-preserved dryopithecin ape from Bavaria. The ulna, femur, tibia, vertebrae, hand and foot bones of this ape reveal unknown aspects of the anatomy of late Miocene apes and enable us to reconstruct what may be the ancestral morphology of the great apes and humans.

Extended limb clambering

The fossils (Fig. 1) include remains of at least four individuals, with a partial skeleton that is sufficiently complete to describe the morphology of the limbs and spine and proportions of the body in detail. The results reveal a combination of anatomical features that are indicative of a pattern of arboreal behaviour that we term extended limb clambering (ELC). It is characterized by generalized limb proportions superimposed on a unique combination of knee, ankle, elbow and wrist postures and strongly grasping extremities. ELC incorporates powerful hallucal grasping, plantigrade feet, extended hip and knees, wide ranging elbow flexion–extension and pronation–supination, a mobile wrist, and hands with curved phalanges and a deep first metacarpal joint. It differs from previously identified forms of positional behaviour. Plantigrade and palmigrade quadrupeds (Old World monkeys and *Ekembo*) lack the suspensory attributes of the forelimb and the extension set of the knee. Knuckle-walkers (chimpanzees, bonobos and gorillas) lack the extended knee and have less powerfully developed hallucal and pollical grasping. The hand phalanges of *Danuvius* also lack the robusticity typical of

[1]Department of Geosciences, Eberhard-Karls-Universität Tübingen, Tübingen, Germany. [2]Senckenberg Centre for Human Evolution and Palaeoenvironment, Tübingen, Germany. [3]National Museum of Natural History, Bulgarian Academy of Sciences, Sofia, Bulgaria. [4]Anatomy & Cell Biology, University of Indianapolis, Indianapolis, IN, USA. [5]Paleontology & Geobiology, Ludwig-Maximilians-Universität, Munich, Germany. [6]Earth Dynamics Research Group, School of Earth and Planetary Sciences, Curtin University, Perth, Western Australia, Australia. [7]Department of Anthropology, University of Toronto, Toronto, Ontario, Canada. *e-mail: m.boehme@ifg.uni-tuebingen.de

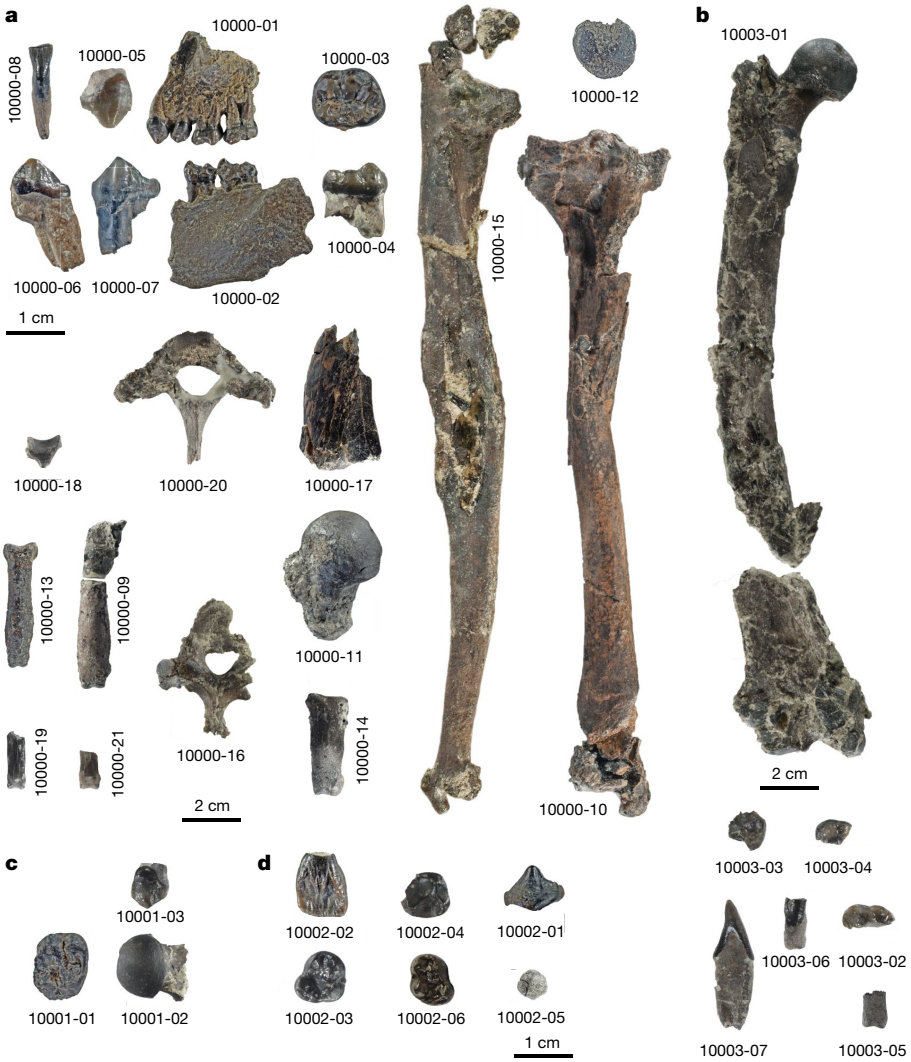


Fig. 1 | Fossil remains of four *D. guggenmosi* individuals from late Miocene sediments of Hammerschmiede. a, Holotype GPIT/MA/10000 male individual. **b**–**d**, Paratype individuals GPIT/MA/10003 (female), GPIT/MA/10001 (female) and GPIT/MA/10002 (juvenile). An excavation plan and a complete list of all elements can be found in Extended Data Fig. 1 and Supplementary Table 2. The scale bar is 20 mm for all bones and 10 mm for all isolated teeth.

knuckle-walkers. Arboreal clambering orangutans lack the weight-bearing adaptations present in the knee and ankle of *Danuvius* and have features that much more strongly emphasize forelimb postural and locomotor adaptations. *Danuvius* is distinguished from all known catarrhines in its vertebral morphology, with an elongated lumbar region combined with spinal invagination/lordosis, which shifts the body mass over the expanded proximal tibial joint surfaces. The uniqueness of ELC is that it does not favour the forelimb or the hindlimb, as in most primates, but utilizes both limbs in roughly equal proportions. ELC includes a combination of joint positions and loading patterns of both hominin bipedalism that emphasize hindlimb extension and spinal curvatures, and extant great ape suspension, which emphasizes powerful and mobile forelimbs. We propose ELC as a new model of the ancestral mode of positional behaviour of the last common ancestor of living great apes and humans. ELC is a precursor to obligate bipedalism, which shifts the emphasis of positional behaviour to the hindlimbs, and to suspension, in which the emphasis shifts to the forelimbs.

Systematic palaeontology

Order Primates Linnaeus, 1758
Infraorder Catarrhini Geoffroy, 1812
Family Hominidae Gray, 1825
Danuvius guggenmosi gen. et sp. nov.

Etymology. The genus name is derived from Celtic–Roman river god Danuvius. The trivial name honours the discoverer of the Hammerschmiede locality, Sigulf Guggenmos.

Holotype. Partial skeleton of male individual GPIT/MA/10000, comprising 21 elements (Fig. 1a): partial left mandible with M_1 and M_2, partial left maxilla with P^3–M^2, isolated mandibular (left I_1, P_3; right P_3, M_2, M_3) and maxillary teeth (right P^3), first and transitional thoracic vertebrae, left humeral shaft fragment, right ulna, left metacarpal I fragment, right proximal manual phalanges II and IV, two left intermediate manual phalanx fragments, right femoral head, right patella, left tibia, left proximal pedal phalanx I.

Paratypes. Two smaller adults (GPIT/MA/10001 (Fig. 1c), comprising left P^3, M^1, left femur head; and GPIT/MA/10003 (Fig. 1b), comprising left I_1, I_2, fragments of M_1, M^1, M^2, left femur, proximal hallucal phalanx fragment) and one juvenile individual (GPIT/MA/10002 (Fig. 1d), comprising unerupted left P_3, left I^1, left and right DP_4, right DP^4, epiphysis of the intermediate manual phalanx).

Locality and horizon. Hammerschmiede Clay pit near Pforzen (Allgäu region, Bavaria, Germany, Extended Data Fig. 1; 47.923° N, 10.588° E); level Hammerschmiede (HAM) 5 at stratigraphic metre 12 in the local section, which has been dated magnetostratigraphically to 11.62 million years ago[15].

Diagnosis. Small hominid ranging in size from about 17 to 31 kg. The palate is narrow and deep with a thick palatine process; the maxilla is high, anteroposteriorly broad, with an anteriorly facing zygomatic root above the distal moiety of P^4, maxillary sinus invaginating the zygomatic and alveolar processes, canine fossa deep and narrow, canine

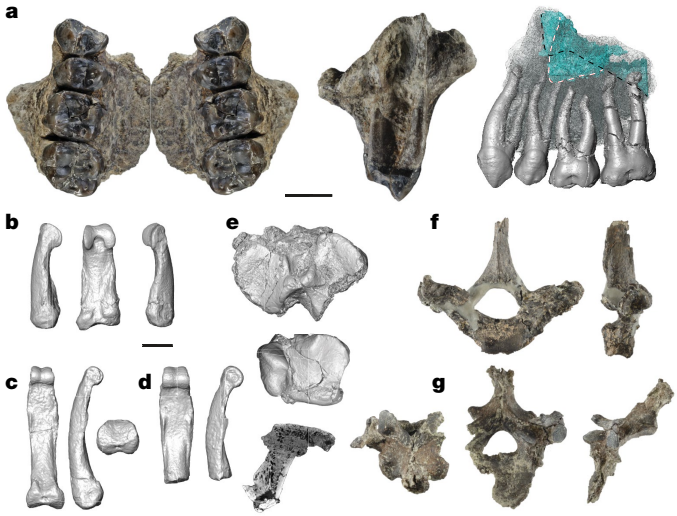


Fig. 2 | *D. guggenmosi* holotype. a, Palate (left; right side mirror-imaged) and left maxilla from superior (middle) and lateral (right) views, with a three-dimensional rendering of dental roots and maxillary sinus (blue). The sinus is invaginated by the posterobuccal and lingual roots of M^2 and is superior to the roots more anteriorly (dashed black line). Laterally the sinus extends into the zygomatic root (dashed white line); additional images are shown in Extended Data Fig. 10. **b**, Left proximal hallucal phalanx in lateral (left), plantar (middle) and medial (right) views. **c**, Right proximal hand phalanx 2 in palmar (left), ulnar (middle) and proximal (right) views. **d**, Right proximal hand phalanx 4 in plantar (left) and ulnar (right) views. **e**, Tibial proximal (top) and distal (middle) articulations (anterior is up) and sagittal computed tomography cross-section through the middle of the lateral condyle (bottom; superior is up). **f**, First thoracic vertebra in superior (left) and left-lateral (right) views. **g**, Diaphragmatic vertebra in posterior (left), superior (middle) and right-lateral (right) views. Scale bars, 10 mm.

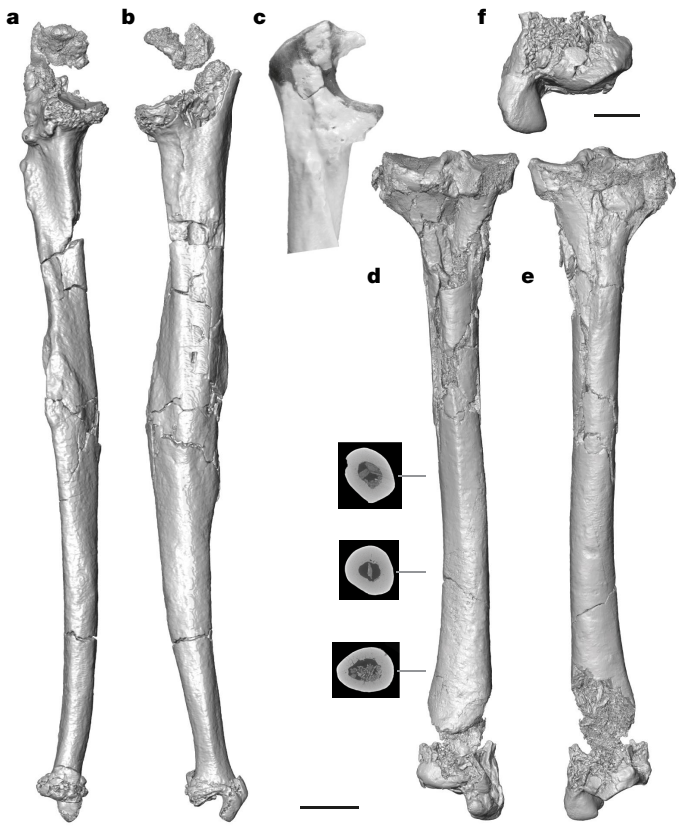


Fig. 3 | *D. guggenmosi*, right ulna (GPIT MA/10000-10) and left tibia (GPIT MA/10000-15). a–**c**, Anterolateral (**a**) and medial (**b**) views of the ulna and the reconstructed proximal end in lateral view (**c**). **d**–**f**, Posterior (**d**) and anterior (**e**) views of the tibia and the distal epiphysis in anterior view (**f**). Tibial shaft cross-sections are given at 20%, 35% and 50% of shaft length from the distal end. Additional images of the ulna and tibia are shown in Extended Data Fig. 4. Scale bars, 20 mm (**a**–**e**) and 10 mm (**f**).

root alveolus vertically oriented; I^1 mesiodistally narrow, high-crowned with a strong lingual pillar and mesial marginal ridge; postcanine dentition with strongly developed crista, P^3 lacks the paraconule, molars are broad relative to the length with compressed trigons and thick enamel; mandibular corpus is low, robust with a prominent mandibular eminence and a broad extramolar sulcus; ulna has a straight shaft, moderately deep proximally, short olecranon, deep, strongly keeled, anteriorly oriented trochlear notch, large, laterally oriented radial notch, large head, short, non-articular styloid process; first metacarpal base strongly dorsopalmarly curved saddle-shaped joint; proximal hand phalanges are long, curved, with strongly developed flexor sheath ridges; femur head projects above the greater trochanter, extension of joint surface onto the superoposterior surface of femoral neck, neck compressed and strongly vertically oriented; tibia with broad proximal end, thickened metaphyses, mediolaterally concave condylar surfaces, lateral condyle anteroposteriorly flat, deeply incised and posteriorly oriented intercondylar notch, prominent intercondylar eminences, trochlear surface roughly square-shaped, strongly keeled, prominent malleolus deeply notched at its base with an anterolaterally expanded joint surface; patella with broad, flat joint surface; proximal hallucal phalanx is large, robust at mid shaft, broad proximally, prominent flexor sheath ridges, strong lateral torsion of the distal end; first thoracic vertebra with short, divergent pedicles, strongly divergent zygapophyseal orientations, univertebral rib articulation; penultimate or antepenultimate diaphragmatic vertebra with a prominent metapophysis.

Differential diagnosis. The craniodental morphology of *Danuvius* is diagnostically dryopithecin ('Expanded differential diagnosis of *D. guggenmosi*' in the Methods). The anterior palate (Fig. 2a) is short in comparison with pongines, with a stepped subnasal fossa, as is typical of dryopithecins and extant hominines. *Danuvius* is distinguished from other dryopithecins in having a unique combination of facial attributes (compressed canine fossa, vertical canine implantation, anteriorly facing malar surface, robust mandible, prominent mandibular eminence, wide extramolar sulcus; Extended Data Fig. 2). The proximal ulna differs from *Hispanopithecus* and *Rudapithecus* in its anteriorly facing trochlear notch and expanded coronoid process (Fig. 3). The distal tibia differs from *Hispanopithecus* in its more squared outline and in details of articular morphology (see Supplementary Information for detailed descriptions and comparisons and Supplementary Tables 3–24 for measurements).

Limb proportions and posture

The postcrania of *Danuvius* reveals numerous previously unknown aspects of dryopithecin morphology. Compared with the length of the tibia, *Danuvius* has a relatively elongated ulna (Fig. 4a and Extended Data Fig. 3), comparable to *Pan paniscus*. In *Pongo*, the ulna is longer whereas in cercopithecoids and early hominins it is shorter. On the basis of reconstructed lengths, *Oreopithecus* and *Hispanopithecus* have tibia:ulna ratios that are comparable to that of *Danuvius*.

A mediolaterally broad thorax and orthogrady is inferred from the dorsal orientation of the thoracic transverse processes, combined with a low costal facet angle on the first thoracic vertebra[16] (Fig. 2f, g). Inferred from the difference in inclination of the spinous processes between the first vertebra and the lower thoracic vertebra, the upper spinal column was substantially curved (cervical lordosis/thoracic kyphosis)[17].

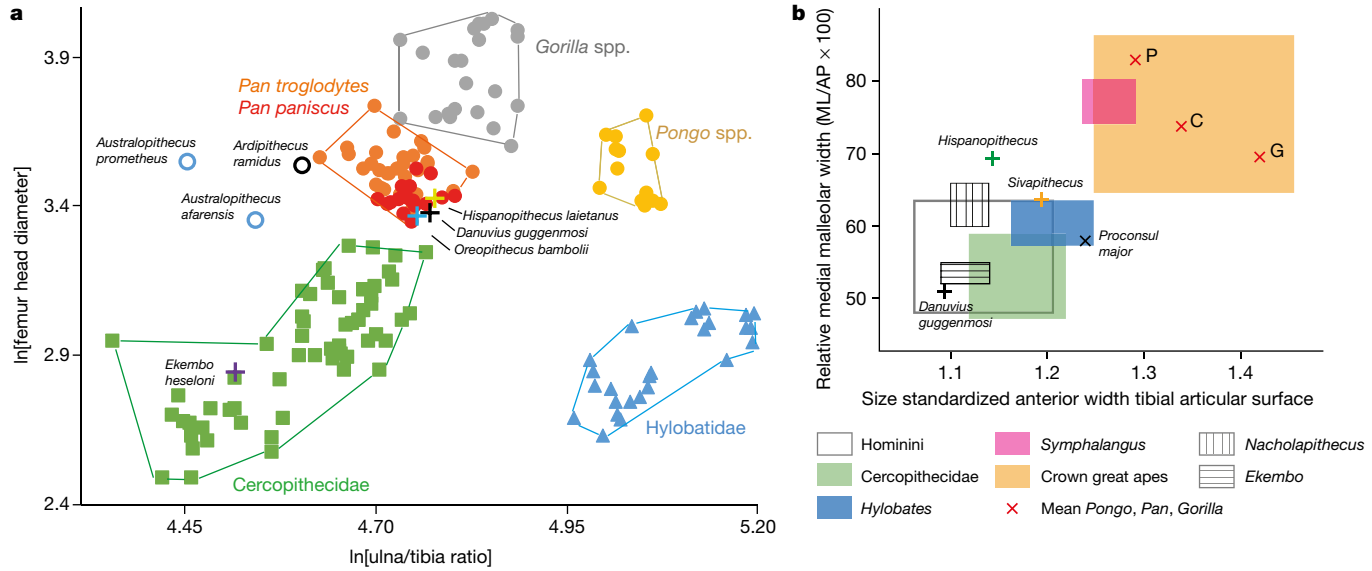


Fig. 4 | Body proportions and distal tibia articulation metrics. a, Ratio of ulna-to-tibia physiologic length (natural logarithm) in relation to body mass (natural logarithm of femur head diameter) of extant catarrhines ($n = 178$; for raw data see Supplementary Table 7) compared to fossil hominoids (*D. guggenmosi*, GPIT/MA10000; *Hispanopithecus laietanus*, IPS 18000; *Oreopithecus bamboli*, IGF 11778; *Ardipithecus ramidus*, ARA-VP-6/500; *Australopithecus prometheus*, StW 573; *Australopithecus afarensis*, A.L. 288-1; *Ekembo heseloni*, KNM-RU 2036; data are from previous studies[9,13,37–40]). **b**, Plot of relative thickness of tibial medial malleolus and size-standardized anterior width of tibial distal articulation surface (measurements follow a previous study[41]) of extant catarrhines, compared to fossil hominoids (*D. guggenmosi*, GPIT/MA/10000; *H. laietanus*, IPS 18000; *Sivapithecus indicus*, YGSP 1656; *Nacholapithecus kerioi*, KNM-BG 35250; *E. heseloni*, KNM-RU 2036, 3589; *Proconsul major*, NAP I'58; comparative data were obtained from previous studies[31,32,41]; for raw data see Supplementary Tables 19, 20). C, *Pan*; G, *Gorilla*; P, *Pongo*.

D. guggenmosi is, to our knowledge, the first Miocene hominid with evidence of diaphragmatic vertebra placement, which is important in interpreting thoracolumbar spine evolution in hominoids[18]. The well-developed costotransversal facet of GPIT/MA/10000-16 (Fig. 2g) indicates a non-ultimate thoracic position for the diaphragmatic vertebra and therefore a functionally longer lower back, as in early hominins, stem-hominoids and cercopithecids[18–24]. On the basis of indirect evidence from the pelvis, a longer lower back has also been inferred for *Rudapithecus*[25]. Extant hominoids including *Homo* show a diaphragmatic placement at the ultimate thoracic vertebra level[24]. The contrasting vertebral configuration of *Danuvius* suggests that diaphragmatic cranial displacement is the symplesiomorphic hominoid condition, supporting the long-back model[26,27]. The increased number of functional lumbar vertebrae allows sagittal flexibility to lordose the lumbar column, which contributes to effectively position the centre of mass over extended hips, knees and plantigrade feet (see below), implying at least some degree of habitual bipedal posture[16].

Positional behaviour

Several skeletal elements of the upper limb bear unmistakable hallmarks of below-branch or suspensory positional behaviour (Fig. 3a–c and Extended Data Fig. 4). Despite the pathology evident on the ulna (Supplementary Information), these include a reduced olecranon process, broad, keeled trochlear notch with prominent medial and lateral surfaces for a trochleaform humeral trochlea, large laterally oriented radial facet, robust proximal ulnar shaft and a reduced, non-articular ulnar styloid process. The proximal hand phalanges are curved with prominent flexor sheath ridges (Fig. 2c, d and Extended Data Figs. 5, 6), indicating that suspension played an important—but not dominant—part in its locomotory repertoire (for example, more similar to *Pan* than to *Pongo*). Powerful pollical grasping and increased thumb mobility are indicated by the strong dorsopalmar and radioulnar curvatures of the base of the first metacarpal (Fig. 1a).

The lower limb suggests postural extension at the hip and knee joints and a uniform force distribution in a stabilized ankle joint, combined with a powerful grasping hallux. On the femur (Fig. 1b and Extended Data Fig. 7b–d), the low greater trochanter, the more vertically oriented neck and the posterosuperior expanded joint surface suggest that the femoral head articulated in habitual extension with an os coxae that was laterally rotated, which would have caused the iliac blade to be more tilted inferolaterally. This may have enhanced the function of the gluteal muscles as hip stabilizers (abductors) in bipedal posture, as in hominins. The flat patella (Fig. 1a and Extended Data Fig. 7a) and shallow rounded patellar surface suggest slow and deliberate movements (Supplementary Information). The absence of an anteroposterior convexity to the lateral tibial condyle (Fig. 2e and Extended Data Fig. 8), a character shared with hominins and hylobatids[28], suggests an extension set to the knee joint, as a flatter contour maximizes tibiofemoral contact area and joint stability during extended knee postures. A buttressing of the tibial metaphysis also reflects stereotypical extended knee postures under compressive load[28,29]. The exceptional development of the intercondylar eminence is probably related to the presence of strongly developed cruciate ligaments. The subequal size of the tibial condyles indicate a more equally distributed weight transmission on the knee joint[30]. Together, the morphology of the tibial plateau suggests an adaptation emphasizing an extended knee reinforced by strongly developed intra-articular ligaments. We interpret the distal tibia of *Danuvius*, with its mediolaterally short anterior trochlear margin and its mediolaterally narrow malleolus (Fig. 4b), to be an adaptation to a more uniform distribution of forces across the joint surface, with limited ankle loading in dorsiflexion and inversion compared to extant apes[31,32]. The combination of the anteroposteriorly deep malleolus, medially expanded joint surface, prominent anterior margin with a strongly developed beak and strongly inclined medial and lateral trochlear surfaces produces a hinge-like morphology to the anterior talocrural joint, which would have been most stable with the foot roughly perpendicular to the long axis of the tibia. This is corroborated by the nearly perpendicularly orientated tibia relative to the horizontal plane of the angle joint (Fig. 3d, e and Supplementary Information). Extant great apes, which load the ankle in inversion during climbing, have an obliquely oriented tibia relative to the plane of the ankle joint[31,33]. The near perpendicular tibial angle is a shared

character between hominins and *Danuvius* and supports the inference of a habitual valgus knee position and bipedalism for the new genus.

A robust, elongated and strongly laterally torsioned hallux (Extended Data Figs. 5b, c, 9) with well-developed muscular attachments suggests an emphasis on powerful hallucal grasping with adducted ankle stabilized in a neutral position relative to the long axis of the tibia. In contrast to extant apes, the hallux was capable of interphalangeal hyperflexion, as indicated by the substantial plantar inter-condylar recess and depression (Fig. 2b), enabling *Danuvius* to securely grasp small-diameter arboreal supports.

Discussion

The uniqueness of *D. guggenmosi* is demonstrated by its small body size (between siamangs and bonobos; Supplementary Information and Supplementary Table 23) with limb proportions most similar to bonobos (Fig. 4a), a cranial shifted diaphragmatic vertebra (Fig. 2g), a strong grasping hallux (Fig. 2b) and a morphology of the tibia that is surprisingly similar to hominins (large-sized and flat lateral condyle with 'buttressed' plateau, tibial shaft perpendicular to talar facet, mediolaterally narrow malleolus and short anterior trochlear margin) (Fig. 3d–f, Extended Data Fig. 4 and Supplementary Information). The combination of morphological attributes of the limbs and vertebra of *Danuvius* point to a newly recognized form of positional behaviour. In contrast to suspensory behaviour, clambering and arm-assisted bipedalism in *Pongo*[34] or climbing and suspension in African apes, ELC involves equal contributions of the fore- and hindlimbs. The foot is flat and adducted on horizontal to mildly inclined branches with a hallux capable of powerful grasping, stabilizing the hindlimb. Torques resulting from body rotation above the knee are countered by powerfully developed cruciate ligaments. The knee is habitually extended and supported by a thickened plateau and large, flat-to-concave, proximally facing condyles. The elbow is capable of a full range of flexion–extension and pronation–supination as in extant hominoids. The hand was strong enough to generate the force to counter torques in a variety of positions ranging from suspensory to palmigrade, but without the hyperextension at the metacarpophalangeal joints that characterize Old World monkeys and *Pierolapithecus*. This newly defined locomotor category includes attributes of orthograde suspension and hominin bipedalism, making it a potential candidate for the positional behaviour of the last common ancestor of great apes and humans. *Danuvius* provides fossil evidence that hominin bipedalism and great ape suspension evolved from a form of arboreal locomotion that incorporates attributes of each[35,36], which has roots in the middle Miocene of Europe.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1731-0.

1. Begun, D. R. in *Biped to Strider: The Emergence of Modern Human Walking* (eds Meldrum, D. J. & Hilton, C. E.) 9–33 (Kluwer, 2004).
2. Begun, D. R. & Kivell, T. L. Knuckle-walking in *Sivapithecus*? The combined effects of homology and homoplasy with possible implications for pongine dispersals. *J. Hum. Evol.* **60**, 158–170 (2011).
3. Richmond, B. G., Begun, D. R. & Strait, D. S. Origin of human bipedalism: the knuckle-walking hypothesis revisited. *Am. J. Phys. Anthropol.* **116**, 70–105 (2001).
4. Crompton, R. H., Sellers, W. I. & Thorpe, S. K. Arboreality, terrestriality and bipedalism. *Phil. Trans. R. Soc. Lond. B* **365**, 3301–3314 (2010).
5. Begun, D. R. Dryopithecins, Darwin, de Bonis, and the European origin of the African apes and human clade. *Geodiversitas* **31**, 789–816 (2009).
6. Begun, D. R., Nargolwalla, M. C. & Kordos, L. European Miocene hominids and the origin of the African ape and human clade. *Evol. Anthropol.* **21**, 10–23 (2012).
7. Alba, D. M. Fossil apes from the vallès-penedès basin. *Evol. Anthropol.* **21**, 254–269 (2012).
8. Langergraber, K. E. et al. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc. Natl Acad. Sci. USA* **109**, 15716–15721 (2012).
9. Moyà-Solà, S. & Köhler, M. A *Dryopithecus* skeleton and the origins of great-ape locomotion. *Nature* **379**, 156–159 (1996).
10. Moyà-Solà, S., Köhler, M., Alba, D. M., Casanovas-Vilar, I. & Galindo, J. *Pierolapithecus catalaunicus*, a new Middle Miocene great ape from Spain. *Science* **306**, 1339–1344 (2004).
11. Alba, D. M., Almécija, S., Casanovas-Vilar, I., Méndez, J. M. & Moyà-Solà, S. A partial skeleton of the fossil great ape *Hispanopithecus laietanus* from Can Feu and the mosaic evolution of crown-hominoid positional behaviors. *PLoS ONE* **7**, e39617 (2012).
12. Begun, D. R. in *Handbook of Paleoanthropology* (eds Henke, W. & Tattersall, I.) 1261–1332 (Springer, 2015).
13. Lovejoy, C. O., Suwa, G., Simpson, S. W., Matternes, J. H. & White, T. D. The great divides: *Ardipithecus ramidus* reveals the postcrania of our last common ancestors with African apes. *Science* **326**, 73–106 (2009).
14. White, T. D., Lovejoy, C. O., Asfaw, B., Carlson, J. P. & Suwa, G. Neither chimpanzee nor human, *Ardipithecus* reveals the surprising ancestry of both. *Proc. Natl Acad. Sci. USA* **112**, 4877–4884 (2015).
15. Kirscher, U. et al. A biochronologic tie-point for the base of the Tortonian stage in European terrestrial settings: magnetostratigraphy of the topmost Upper Freshwater Molasse sediments of the North Alpine Foreland Basin in Bavaria (Germany). *Newsl. Stratigr.* **49**, 445–467 (2016).
16. Williams, S. A. & Russo, G. A. Evolution of the hominoid vertebral column: the long and the short of it. *Evol. Anthropol.* **24**, 15–32 (2015).
17. Latimer, B. & Ward, C. V. in *The Nariokotome* Homo erectus *Skeleton* (eds Walker, A. & Leakey, R.) 266–293 (Springer, 1993).
18. Williams, S. A., Middleton, E. R., Villamil, C. I. & Shattuck, M. R. Vertebral numbers and human evolution. *Am. J. Phys. Anthropol.* **159**, 19–36 (2016).
19. Haeusler, M., Regula, S. & Thomas, B. Modern or distinct axial bauplan in early hominins? A reply to Williams (2012). *J. Hum. Evol.* **63**, 557–559 (2012).
20. Nakatsukasa, M. & Kunimatsu, Y. *Nacholapithecus* and its importance for understanding hominoid evolution. *Evol. Anthropol.* **18**, 103–119 (2009).
21. Pilbeam, D. The anthropoid postcranial axial skeleton: comments on development, variation, and evolution. *J. Exp. Zool.* **302B**, 241–267 (2004).
22. Ward, C. V., Walker, A., Teaford, M. F. & Odhiambo, I. Partial skeleton of *Proconsul nyanzae* from Mfangano island, Kenya. *Am. J. Phys. Anthropol.* **90**, 77–111 (1993).
23. Ward, C. V., Nalley, T. K., Spoor, F., Tafforeau, P. & Alemseged, Z. Thoracic vertebral count and thoracolumbar transition in *Australopithecus afarensis*. *Proc. Natl Acad. Sci. USA* **114**, 6000–6004 (2017).
24. Williams, S. A. Placement of the diaphragmatic vertebra in catarrhines: implications for the evolution of dorsostability in hominoids and bipedalism in hominins. *Am. J. Phys. Anthropol.* **148**, 111–122 (2012).
25. Ward, C. V., Hammond, A. S., Plavcan, J. M. & Begune D. R. A late Miocene hominid partial pelvis from Hungary. *J. Hum. Evol.* https://doi.org/10.1016/j.jhevol.2019.102645 (2019).
26. McCollum, M. A., Rosenman, B. A., Suwa, G., Meindl, R. S. & Lovejoy, C. O. The vertebral formula of the last common ancestor of African apes and humans. *J. Exp. Zool.* **314B**, 123–134 (2010).
27. Lovejoy, C. O. & McCollum, M. A. Spinopelvic pathways to bipedality: why no hominids ever relied on a bent-hip–bent-knee gait. *Phil. Trans. R. Soc. Lond. B* **365**, 3289–3299 (2010).
28. Landis, E. K. & Karnick, P. A three-dimensional analysis of the geometry and curvature of the proximal tibial articular surface of hominoids. In *Proc. SPIE 60560 Three-Dimensional Image Capture and Applications VII 60560K* (International Society for Optics and Photonics, 2006).
29. Frelat, M. A. et al. Evolution of the hominin knee and ankle. *J. Hum. Evol.* **108**, 147–160 (2017).
30. Tardieu, C. Ontogeny and phylogeny of femoro-tibial characters in humans and hominid fossils: functional influence and genetic determinism. *Am. J. Phys. Anthropol.* **110**, 365–377 (1999).
31. DeSilva, J. M. Functional morphology of the ankle and the likelihood of climbing in early hominins. *Proc. Natl Acad. Sci. USA* **106**, 6567–6572 (2009).
32. DeSilva, J. M., Morgan, M. E., Barry, J. C. & Pilbeam, D. A hominoid distal tibia from the Miocene of Pakistan. *J. Hum. Evol.* **58**, 147–154 (2010).
33. Latimer, B., Ohman, J. C. & Lovejoy, C. O. Talocrural joint in African hominoids: implications for *Australopithecus afarensis*. *Am. J. Phys. Anthropol.* **74**, 155–175 (1987).
34. Thorpe, S. K., Holder, R. L. & Crompton, R. H. Origin of human bipedalism as an adaptation for locomotion on flexible branches. *Science* **316**, 1328–1331 (2007).
35. Thorpe, S. K., McClymont, J. M. & Crompton, R. H. The arboreal origins of human bipedalism. *Antiquity* **88**, 906–914 (2014).
36. Wolpoff, M. *Australopithecus*: a new look at an old ancestor (part 2).*Gen. Anthropol.* **3**, 1–5 (1997).
37. Straus, W. in *Classification and Human Evolution* (ed. Washburn, S. L.) 146–177 (Aldine, 1963).
38. Asfaw, B. et al. *Australopithecus garhi*: a new species of early hominid from Ethiopia. *Science* **284**, 629–635 (1999).
39. Ruff, C. B. Long bone articular and diaphyseal structure in Old World monkeys and apes. II: estimation of body mass. *Am. J. Phys. Anthropol.* **120**, 16–37 (2003).
40. Haile-Selassie, Y. et al. An early *Australopithecus afarensis* postcranium from Woranso-Mille, Ethiopia. *Proc. Natl Acad. Sci. USA* **107**, 12121–12126 (2010).
41. DeSilva, J. M. *Vertical Climbing Adaptations in the Anthropoid Ankle and Midfoot: Implications for Locomotion in Miocene Catarrhines and Plio-Pleistocene Hominins*. PhD thesis, Univ. Michigan, (2008).

Methods

Geology, age, fossils and taphonomy

The HAM 5 channel represents a riffle pool sequence of a small and shallow meandering rivulet with a talweg width of 4–5 m and a maximum pool depth of 1 m. The gravelly bed load is composed exclusively of reworked pedogenic carbonate concretions that are typically 4–8 mm in diameter. Similar concretions are abundant in Bk palaeosol horizons of the bedrock, indicating a local source of HAM 5 rivulet. Magnetostratigraphy of the local 26-m thick section, combined with a nearby 150 m deep drill core, revealed the date of the channel fill of 11.620 million years ago (±5 thousand years), directly at the base of the Tortonian, late Miocene[15]. Excavation of about 200 m^2 between 2011 and 2018 revealed a high vertebrate diversity that comprised 100 species of fishes, amphibians, reptiles, birds and mammals (see Supplementary Table 1 for faunal list). Hominids are a common element in this thanatocoenosis, representing about 10% of all excavated large mammal individuals. Excavation demonstrates that fossil vertebrates are found exclusively along the channel, suggesting some sort of accumulation. Most finds are disarticulated skeletal elements, which tend to be complete in small- and medium-sized mammals (for example, carnivores, artiodactyls and primates) and broken and sometimes abraded in large-size taxa (for example, perissodactyls and proboscideans). Skeletal articulation occurs in rare cases. However, many medium-sized individuals are documented by associated specimens found within a few square metres, suggesting minor transport and sorting of bones. The 21 bones and teeth from the most complete hominid individual GPIT/MA/10000 represent about 15% of the skeleton. It is found within the talweg at a maximum distance of 20 m, except the first thoracic vertebra, which was found a further 10 m downstream. Moderate sorting of GPIT/MA/10000 is documented by proximal concentration of isolated teeth, followed by skull elements and more distally long bones and phalanges, whereas vertebra are transported furthest down the channel (Extended Data Fig. 1). This arrangement follows experimentally observed patterns of bone taphocoenosis in rivers[42].

Fossil repository

All Hammerschmiede fossils are stored in the palaeontological collection of the University of Tübingen (acronym GPIT), a research infrastructure of the Senckenberg Institute for Human Evolution and Palaeoenvironment (SHEP) Tübingen.

Bone preservation

The Hammerschmiede locality is an active clay-mining pit. Sediments from the fossiliferous rivulet channel HAM 5 are composed of fine-pebbly pedogenic carbonate nodules and marls with various degrees of silt and rare fine-sand admixture. Owing to mining activities, water-saturated clay-rich sediments on steep section walls tend to creep and heavy machinery add compressive load on the sediment surface. Therefore, postcranial long bones of smaller large mammals (for example, deer, tragulids, carnivores and primates) tend to be compressed at the shaft and occasionally laterally distorted. This strongly affected the complete femur of GPIT/MA/10003 (shaft compressed by machinery loading, folded along the shaft due to ground creeping), which was embedded in soft clay. The complete ulna of GPIT/MA/10000 is uncompressed, but at midshaft the cortical bone of the down-lying side is crushed and pushed into the shaft, probably by load compression. Computer tomographic images show that this preservation was facilitated by midshaft osteoporosis. By contrast, the complete tibia of GPIT/MA/10000, embedded in a less compressible silt-dominated matrix, is not crushed along the shaft, but laterally distorted at the tuberosity and slightly damaged at medial condyle and distal metaphysis, which are the result of excavation artefacts. Importantly, all cranial and small postcranial ape specimens (phalanges, metapodial, carpal bone and patella), as well as long-bone joint articulations remained undisturbed, but occasionally show small excavation artefacts.

Length reconstruction

To measure the total and physiologic length of distorted long bones, we use three-dimensional prints of virtual reconstructions for the holotype tibia and ulna (GPIT/MA/10000-10 and -15, respectively). The total length of the crushed paratype femur (GPIT/MA/10003-01) is estimated with an accuracy of about ±5 mm.

Expanded differential diagnosis of *D. guggenmosi*

The molars lack cingula and are elongated relative to length, with peripheralized cusps. These attributes and P_3 cusp morphology, P_4 length and M_1–M_2 proportions distinguish *Danuvius* from *Ekembo* and other early Miocene hominoids. The dentition is readily distinguished from thickly enamelled middle and late Miocene apes such as *Kenyapithecus*, *Nacholapithecus*, *Griphopithecus*, *Sivapithecus* and *Ouranopithecus*.

The maxilla of *D. guggenmosi* (Figs. 1a, 2a and Extended Data Figs. 2a, 3a) differs from *Anoiapithecus*, *Pierolapithecus* and *Dryopithecus* in its anteroposteriorly broad zygomatic root (zygomatico-alveolar crest) and convex and postero-inferiorly inclined temporal surface; deeply invaginated maxillary sinus floor; vertically implanted upper male canine (supero-inferiorly and mediolaterally); deep, anteropostreriorly narrow canine fossa and anteriorly facing zygoma. Differs from *Hispanopithecus* and *Rudapithecus* maxilla in its deep, anteropostreriorly narrow canine fossa and anteriorly facing zygoma, anteriorly positioned zygomaticoalveolar crest and deeper palate. Maxillary dentition differs from *Anoiapithecus*, *Pierolapithecus* and *Dryopithecus* by broader premolars; triangular P^3; low mesial and distal P^3 buccal shoulders; more mesiodistally centralized premolar cusps (shorter talon); broad, concave premolar trigon and talon basins; more strongly developed molar crista; more peripheralized cusps; mesiodistally compressed trigon. I^1 differs from *Pierolapithecus* and cf. *Dryopithecus* sp. (La Grive) in its more strongly developed mesial marginal ridge and convex lingual surface. The maxillary dentition differs from *Hispanopithecus* and *Rudapithecus* in its low P^3 crown shoulders and broad upper premolars. The mandible (Fig. 1a and Extended Data Fig. 2b) differs from *Anoiapithecus* and *Dryopithecus* in its shallower, robust corpus (unknown in *Pierolapithecus*), prominent mandibular eminence and wide extramolar sulcus. Mandibular dentition differs from *Anoiapithecus* and *Dryopithecus* in its lower crowned, mesially more vertical P_3 with a prominent mesial beak; broader molar trigonid and talonid basins; shorter mesial fovea; absence of buccal cingula; elongated molars; short M_1 roots (not visible in *Anoiapithecus*). The mandible differs from *Hispanopithecus* and *Rudapithecus* mandibles in the same way as from *Anoiapithecus*, *Pierolapithecus* and *Dryopithecus* and from the lower teeth of *Hispanopithecus* and *Rudapithecus* in having restricted mesial and distal fovea. The mandibular dentition differs from *Ouranopithecus* as it is smaller with more thinly enamelled teeth and it differs in other attributes as in *Rudapithecus* and *Hispanopithecus*. It also differs from *Oreopithecus* in having lower postcanine cusps, less strongly developed crista/cristids, no centroconid, higher P_4 talonid, higher crowned I^1, no upper postcanine lingual cingula. The maxilla differs from early and middle Miocene hominoids in the high position of the zygomatic root. The dentition differs from early and middle Miocene hominoids in the absence of molar cingula, first and second molars of similar size, peripheralized molar cusps, more vertical mesiobuccal P^3 surface and short P^4 shoulders, and higher P_4 talonid.

The partial skeleton GPIT/MA/10000 includes dental and postcranial remains that are much larger than the other Hammerschmiede individuals. This along with the strongly flared mesiobuccal face of the P_3 (Fig. 1a and Extended Data Fig. 2g, j) and the large, elongated canine alveolus (Fig. 2a) strongly imply that GPIT/MA 10000 is a male.

Body mass calculations

For the calculation of the body mass of the individuals, we used metric traits (individual measurements) from hind limbs (femur and tibia) because they are most involved in weight carrying during locomotion in great apes[39]. Our univariate body-mass predictions are based on regression equations from a previously published study[39] for sex/species means of hominoids. In addition, as we can show that body proportions of the male individual GPIT/MA/10000 fall within the range of bonobos and chimpanzees, we assume a comparable scaling pattern and apply regression equations established previously[43] for femur head size of the genus *Pan*. Both methods produce very similar results for the male individual within the 50% confidence interval (Supplementary Table 23). Femur size of the two female specimens GPIT/MA/10001 and GPIT/MA/10003 are significantly lower than of any extant great ape, and hence outside any hominid comparative sample. We therefore use the previously compiled regression equations[39] for the total primate sample (hominoids plus cercopithecoids) for the predictor femur head size and cercopithecoid equations for predictions based on femoral condyle breadth (as recommended in the previously published study[39]).

Calculations of enamel thickness

We used the right M_2 of the holotype (GPIT/MA/10000-03) to calculate enamel thickness given its low occlusal wear (slightly higher on mesial half, wear stage 1–2 according to a previously published study[44]). This tooth was scanned with a FF35 CT at the YXLON Application centre in Heilbronn (Germany) and captured at 170 kV and 55 µA (500-ms exposure time), obtaining a voxel size of 15.8 µm (Extended Data Fig. 3b). Following a previously published study[45], virtual buccolingual sections of the molar were performed using Avizo 9.0. Mesial and distal virtual sections were defined by the tips of the metaconid–protoconid and entoconid–hypoconid perpendicular to the cervical plane. The following variables were measured two-dimensionally in both planes: dentine area (*b*), enamel cap area (*c*), length of the enamel–dentine junction (*e*) and the bi-cervical diameter. The average enamel thickness was calculated as *c*/*e* and the relative enamel thickness (RET) was calculated as previously described[46] using $\mathrm{RET} = 100 \times \left[\frac{\left(\frac{c}{e}\right)}{\sqrt{b}}\right]$. For GPIT/MA/10000-03, the RET = 19.36, based on data from the least worn distal section (Supplementary Table 6).

Ellipse estimates of lateral tibial condyle curvature

To estimate the shape of the lateral tibial condyle, we performed a cut through the sagittal mid-line of the condyle on the three-dimensional scans of tibiae from *D. guggenmosi* (Fig. 2e) and extant catarrhines (Extended Data Fig. 6) using an Artec Space Spider with Artec Studio 11 (three-dimensional scans) and Avizo 9 (cross-sections). Subsequently, the cross-sections were digitalized and a best-fit ellipse was obtained using a non-iterative MATLAB function ('EllipseDirectFit'; from N. Chernov (code available from https://www.mathworks.com/)). To compare the individual ellipses, we further calculated the eccentricity $e = \sqrt{(1 - (b/a)^2)}$ in which *a* and *b* are the semi-major and semi-minor axes with $a \geq b$. The closer *e* is to 1, the more elongated the ellipse is, whereas $e = 0$ represents a circle.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data generated or analysed during this study are included in this published Article (and its Supplementary Information). The computed tomography scans are available from the corresponding author on reasonable request. The new taxon has the following Life Science Identifier: http://zoobank.org/References/E1573024-9543-4B1E-A79B-6E40896A4617.

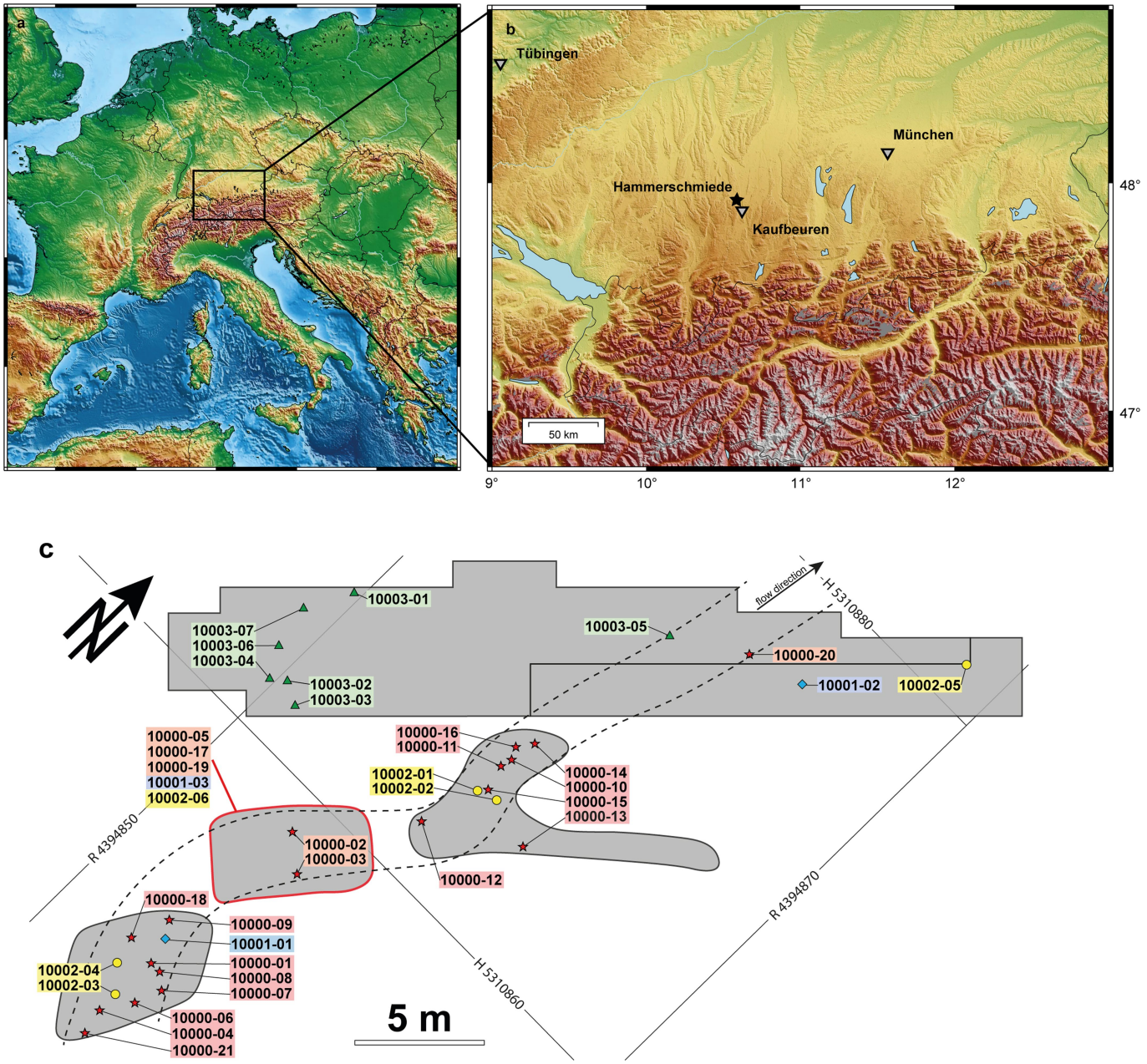
42. Behrensmeyer, A. K. The taphonomy and paleoecology of Plio-Pleistocene vertebrate assemblages east of Lake Rodolf, *Kenya. Bull. Mus. Comp. Zool*. **146**, 473–578 (1975).
43. Almécija, S. et al. The femur of *Orrorin tugenensis* exhibits morphometric affinities with both Miocene apes and later hominins. *Nat. Commun*. **4**, 2888 (2013).
44. Smith, B. H. Patterns of molar wear in hunger-gatherers and agriculturalists. *Am. J. Phys. Anthropol*. **63**, 39–56 (1984).
45. Smith, T. M., Olejniczak, A. J., Martin, L. B. & Reid, D. J. Variation in hominoid molar enamel thickness. *J. Hum. Evol*. **48**, 575–592 (2005).
46. Martin, L. B. *The Relationships of the later Miocene Hominoidea*. PhD thesis, Univ. College London (1983).
47. Wessel, P. et al. Generic mapping tools: improved version released. *EOS* **94**, 409–410 (2013).
48. Amante, C. & Eakins, B. W. *ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis NOAA Technical Memorandum NESDIS NGDC-24* https://www.ngdc.noaa.gov/mgg/global/relief/ETOPO1/docs/ETOPO1.pdf (National Geophysical Data Center, NOAA, 2009).
49. Reuter H. I., Nelson, A. & Jarvis, A. An evaluation of void-filling interpolation methods for SRTM data. *Int. J. Geogr. Inf. Sci*. **21**, 983–1008 (2007).
50. Almécija, S., Alba, D. M. & Moyà-Solà, S. *Pierolapithecus* and the functional morphology of Miocene ape hand phalanges: paleobiological and evolutionary implications. *J. Hum. Evol*. **57**, 284–297 (2009).

Acknowledgements We are indebted to the following researchers and curators for granting access to collections under their care: S. Moyà-Solà and D. Alba, E. Gilissen, L. Costeur, A. van Heteren, S. Merker, E. Weber. We thank C. Schulbert and J.-F. Metayer for computed tomography scanning, and A. Fatz, H. Stöhr and W. Gerber for technical support.

Author contributions M.B. and D.R.B. designed the study; M.B., N.S., J.F., A.T., A.S.D., J.P., U.K., T.L. and D.R.B. collected the data and performed the analyses; M.B., D.R.B. and N.S. discussed the results and wrote the paper.

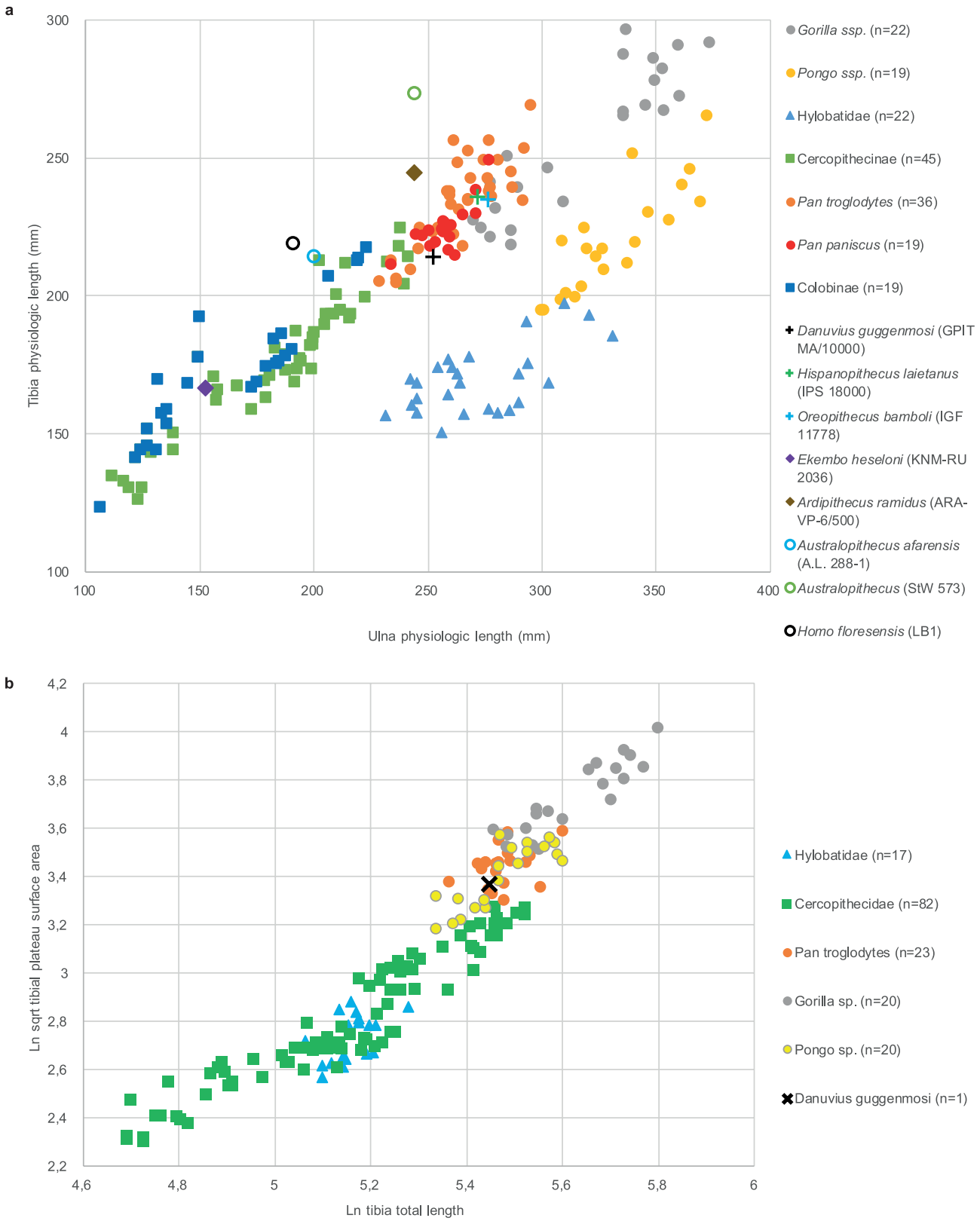
Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1731-0.
Correspondence and requests for materials should be addressed to M.B.
Peer review information *Nature* thanks Jeremy M. DeSilva, Tracy Kivell and Salvador Moyà-Solà for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



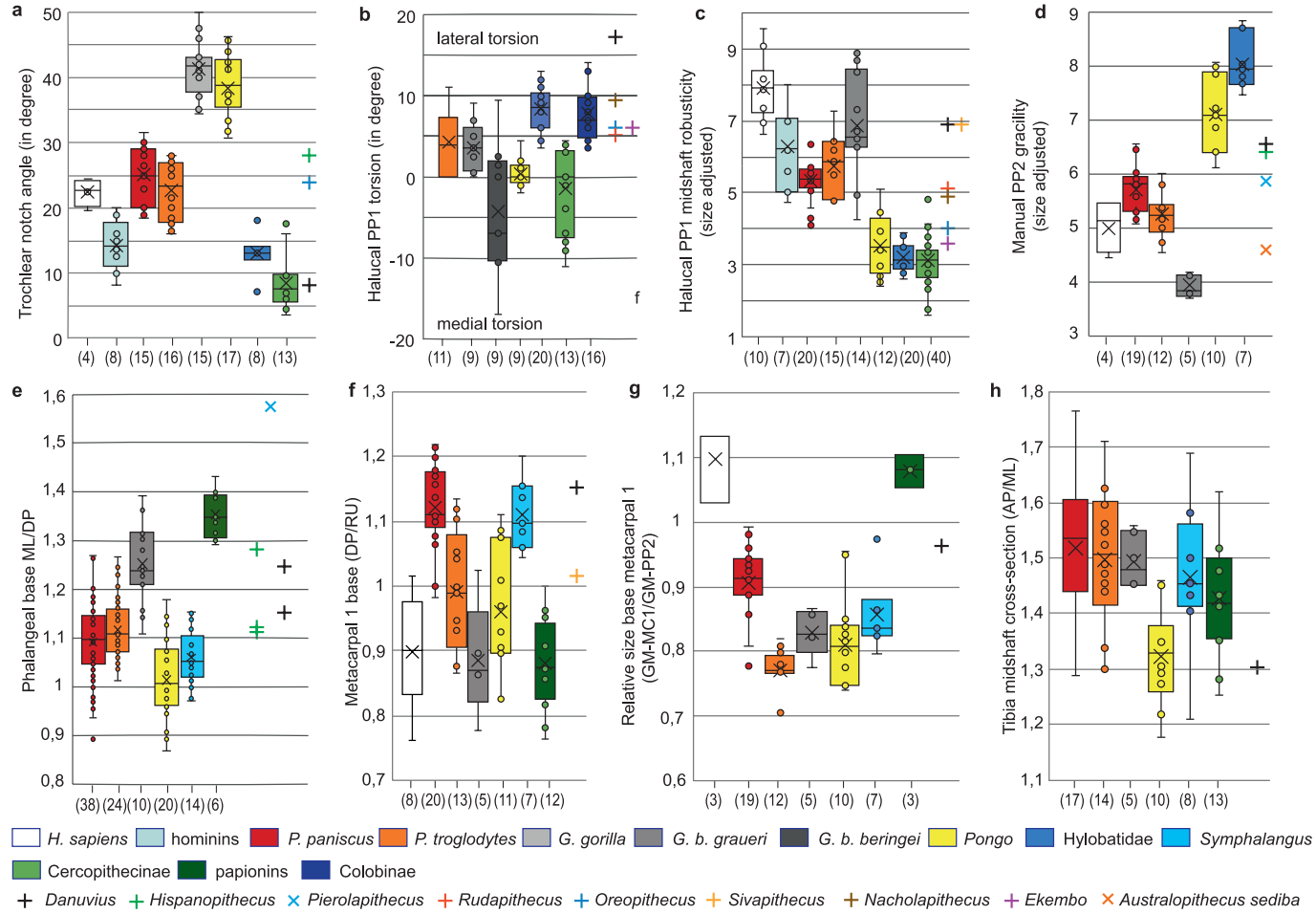
Extended Data Fig. 1 | Localization of Hammerschmiede locality and excavation plan with localized *D. guggenmosi* specimens. a, Topographical map of Europe. **b**, Magnification of the western part of the south German Molasse Basin (North Alpine Foreland Basin). The Hammerschmiede locality (47° 55′ 37″ N, 10° 35.5′ E) is highlighted with a black star. Both maps were created using Generic Mapping Tools[47] and topographic datasets ETOPO1[48] and SRTM3[49]. **c**, Excavation plan of the HAM 5 layer (the section has previously been published[15]) with excavated areas coloured in grey. Intermediate regions represent material lost due to clay mining. Dashed lines indicate the reconstructed thalweg course of the palaeochannel. Different colours and symbols indicate the individual context: holotype (GPIT/MA/10000) adult male marked in red (stars), paratype (GPIT/MA/10001) female 1 in blue (diamonds), paratype (GPIT/MA/10002) juvenile individual in yellow (circles) and paratype (GPIT/MA/10003) female 2 in green (triangles). The red encircled sector indicates removed and stored sediments that were screen washed separately. This area was under threat of destruction from quarry activity. To avoid the complete loss of this sediment, approximately 25 tonnes were removed for remote processing. Two specimens were recovered in situ in this area. Five other specimens from this area were recovered during subsequent screen washing and cannot be more precisely localized. Coordinates correspond to Gauss-Krüger Zone 4 grid with easting (R) and northing (H) in metres.

Extended Data Fig. 2 | *D. guggenmosi*, dental and cranial specimens. a, Left maxilla with P^3–M^2 (GPIT MA/10000-01) in lateral, anterior, medial (top), palatal, posterior, superior (bottom) views. **b**, Left mandible (GPIT MA/10000-02) in lateral, anterior, medial and occlusal views. **c**, Left upper central incisor (GPIT MA/10002-01) in labial, lingual and occlusal views. **d**, Right upper P^3 fragment (GPIT MA/10000-05) in buccal, occlusal and mesial views. **e**, Left P^3 (GPIT MA/10001-03) in buccal, occlusal and mesial views. **f**, Right upper M^1 (GPIT MA/10001-01) in occlusal, medial, distal and buccal views. **g**, Left lower P_3 (GPIT MA/10000-07) in medial, buccal, lingual and occlusal views. **h**, Left lower lateral incisor (GPIT MA/10003-5) in distal, mesial, lingual and labial views. **i**, Left lower central incisor (GPIT MA/10000-08) in distal, mesial and lingual views. **j**, Right lower P_3 (GPIT MA/10000-06) in mesial, distal, buccal and occlusal views. **k**, Right lower M_2 (GPIT MA/10000-03) in lingual, buccal (top), mesial, distal (bottom) and occlusal views. **l**, Right lower M_3 (GIPT MA/10000-04) in lingual, mesial (top), buccal, distal (bottom) and occlusal views. Scale bar, 10 mm.

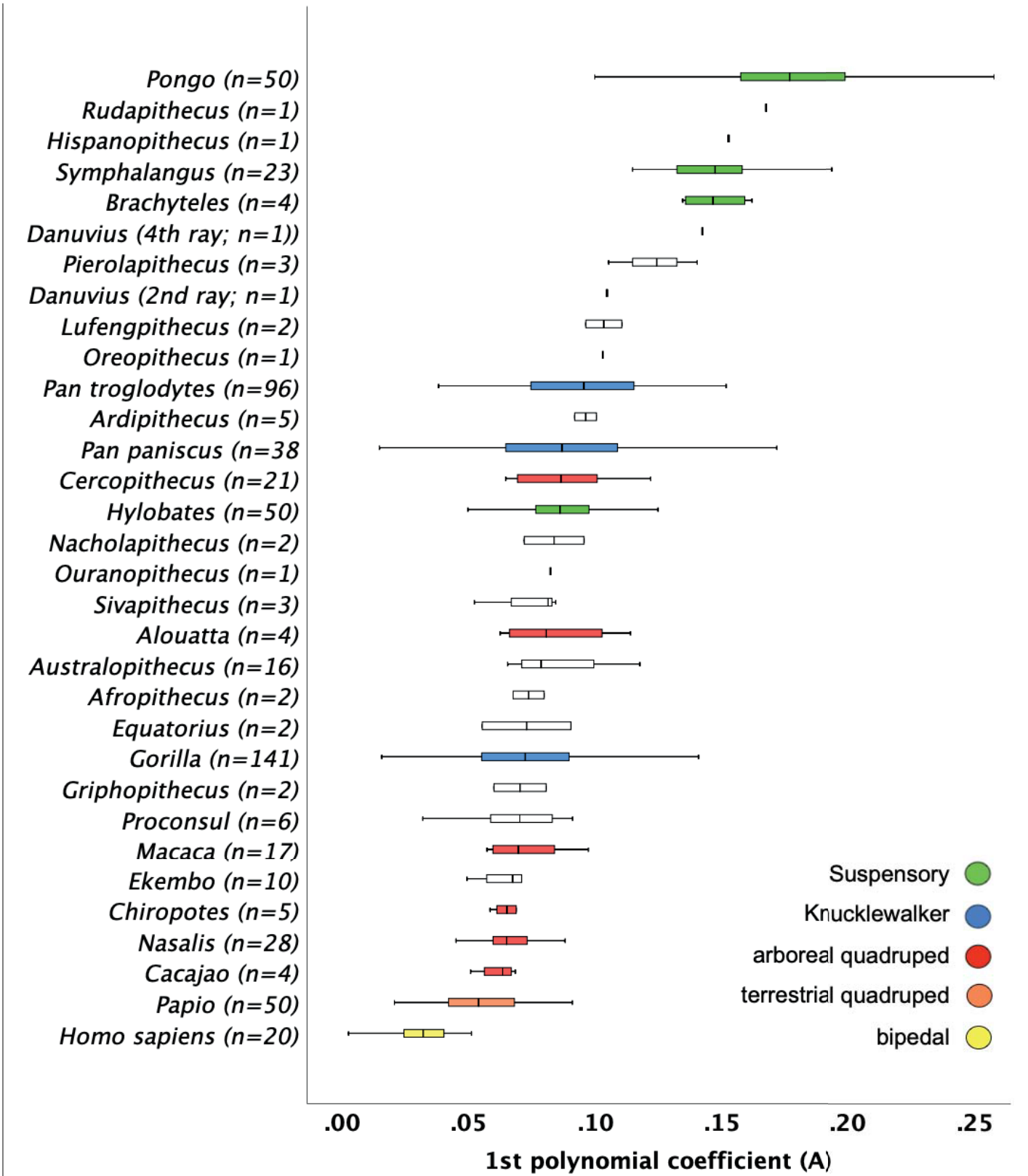


Extended Data Fig. 3 | Long-bone relationships and tibial plateau surface area. a, Relationships of physiologic lengths of tibia and ulna among extant and fossil catarrhines. **b**, Relationships of tibial plateau surface area (TPSA sensu[39], natural logarithm of square root) and tibial total length (natural logarithm) among extant hominids, hylobatids and cercopithecids (comparative data from a previous study[39]). The tibial plateau surface area of GPIT MA/10000-10 is 1,457 mm^2.

Extended Data Fig. 4 | *D. guggenmosi*, additional views of right ulna (GPIT MA/10000-10) and left tibia (GPIT MA/10000-15). a–**d**, Lateral (**a**), anteromedial (**b**) and posterior (**c**) views of the ulna and the reconstructed olecranon in anterior view (**d**). **e**, **f**, Medial (**e**) and lateral (**f**) views of the tibia. Scale bar, 20 mm.

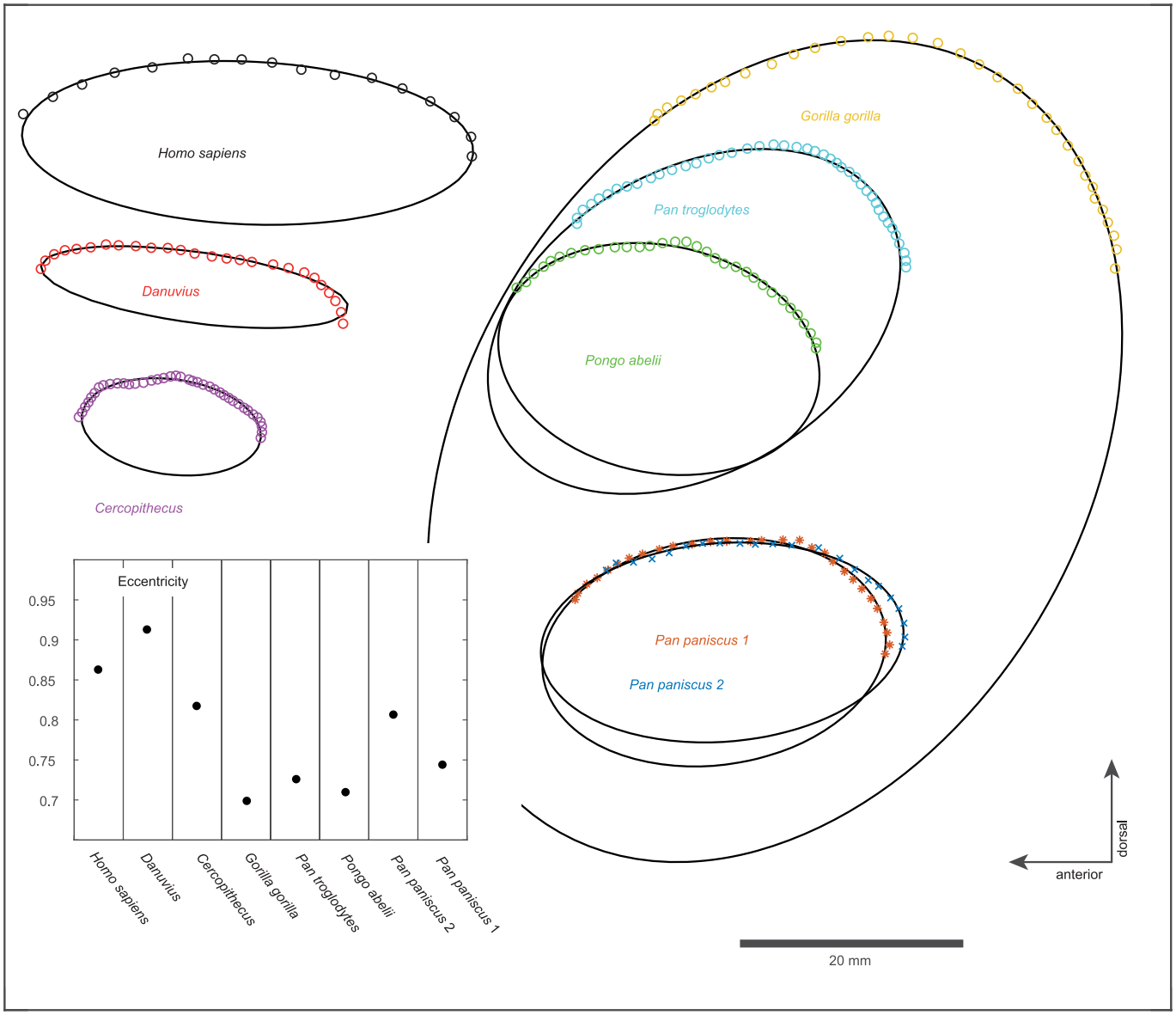


Extended Data Fig. 5 | Ulnar trochlear notch, phalangeal, metacarpal and tibial midshaft comparisons. a, Ulnar trochlear notch angle (for raw data, see Supplementary Table 9). **b**, Hallucal proximal phalanx (PP1) torsion (for measurement, see Methods; for raw data, see Supplementary Table 23). **c**, Size-adjusted hallucal proximal phalanx (PP1) midshaft robusticity (MLms × DPms/GM in which MLms is the mediolateral width at midshaft, DPms is the dorsopalmar height at midshaft and GM is the geometric mean of the seven measurements: ML and DP at proximal, distal and midshaft, and total length; for raw data, see Supplementary Table 22). **d**, Size-adjusted second manual proximal phalanx (PP2) gracility (TL/GM in which TL is the total length and GM is the geometric mean of five measurements: ML and DP at distal and midshaft, and TL; five measurements are used to include *Pierolapithecus catalaunicus*, in which the proximal articulation is damaged[50]; for raw data, see Supplementary Table 11). **e**, Manual phalangeal base, ratio of mediolateral (ML) to dorsopalmar (DP) length (for raw data, see Supplementary Tables 11, 12). **f**, Manual metacarpal 1 base, ratio of dorsopalmar to radioulnar (RU) length (for raw data, see Supplementary Table 10). **g**, Relative size of manual metacarpal 1 base (geometric mean of dorsopalmar and radioulnar lengths) to proximal phalanx of ray 2 (geometric mean of seven measurements; for raw data, see Supplementary Tables 10, 11). **h**, Tibial cross-section at midshaft (ratio of anteroposterior and mediolateral width; for raw data see Supplementary Table 21). Sample sizes (*n*) of biologically independent animals are reported in parentheses below each box plot. All box plots show the centre line (median), box limits (upper and lower quartiles), crosses (arithmetic mean), whiskers (range) and individual values (circles).

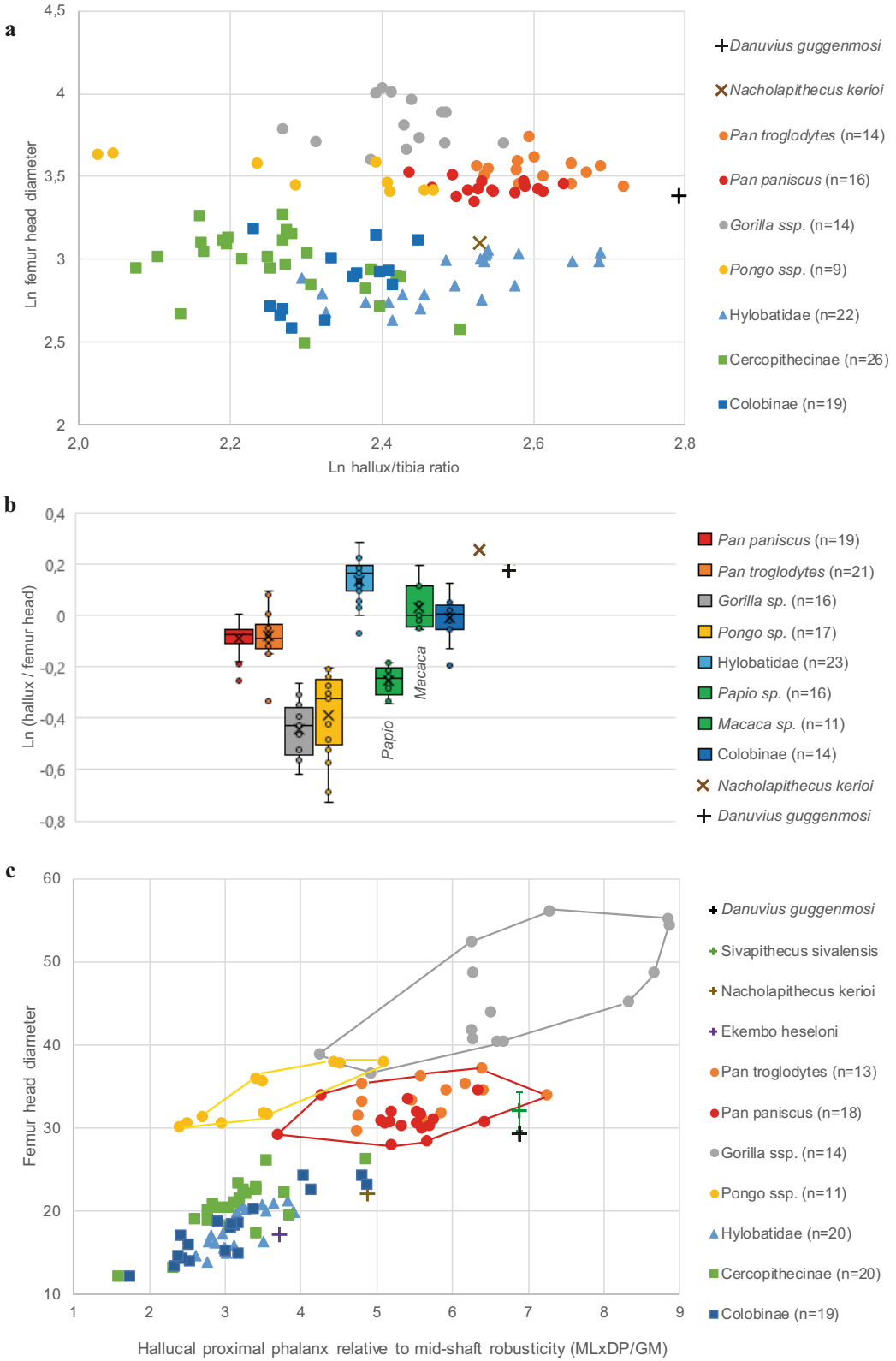


Extended Data Fig. 6 | Curvature manual proximal phalanges. Box plots of the first polynomial coefficient (A) of the second-order polynomial functional representing phalangeal shaft curvature. The box represents the interquartile range, which represents 50% of the sample values. The whiskers are lines that extend from the interquartile range box to the highest and lowest values, excluding outliers. The line across the box indicates the median sample value for coefficient A. Extant primates are colour-coded according to locomotor adaptation. Taxa are arranged according to ascending median phalangeal shaft curvature. Sample sizes (*n*) of biologically independent animals are reported in parentheses after the species names.

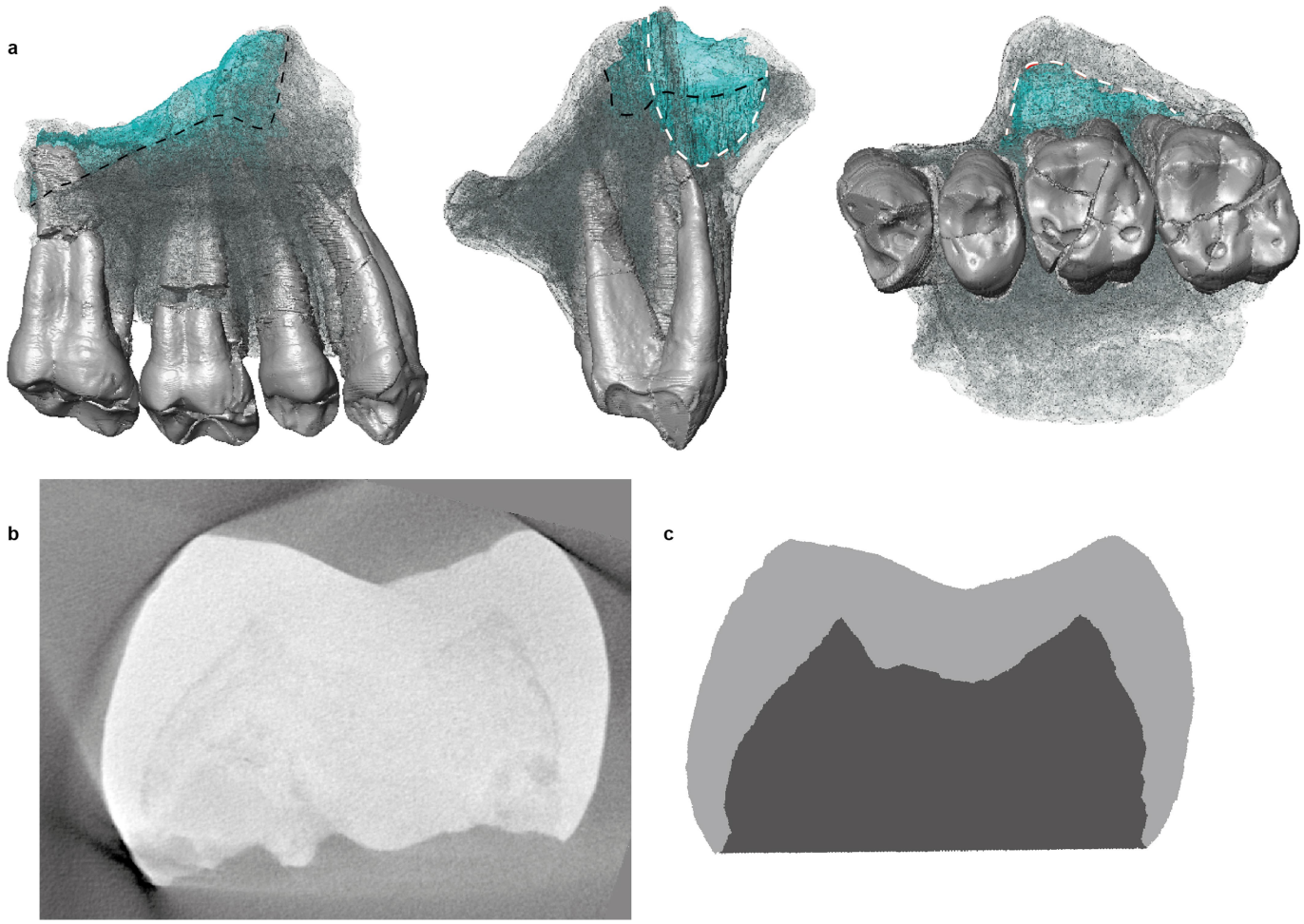
Extended Data Fig. 7 | *D. guggenmosi*, patella and femora. a, Right patella (GPIT MA/10000-12) in external and internal views. **b**, Right femur head (GPIT MA/10000-11) in medial, anterior, posterior (top), superior and lateral (bottom) views. **c**, Left femur head (GPIT MA/10001-02) in medial, posterior, anterior (top), superior and lateral (bottom) views. **d**, Left femur, proximal half (GPIT MA/10003-01) in anterior (top) and posterior (bottom) views. Scale bar, 10 mm.



Extended Data Fig. 8 | Ellipse estimates of lateral tibial condyle. Best fit ellipses to digitalized portions of sagittal cross-sections through lateral tibial condyle of *D. guggenmosi* and extant catarrhines. Digitalized dots are shown in colour and best-fit ellipses in black. Orientation of ellipses follows the lateral condyle orientation (dorsal is up, anterior is left) at the same scale (scale bar, 20 mm). Inset shows calculated values of eccentricity for the obtained ellipses. Results indicate that both *Danuvius* and extant humans have a flat lateral tibial condyle (eccentricity >0.85), whereas great apes exhibit a convex lateral condyle (eccentricity <0.80) and *Cercopithecus* occupy an intermediate position.



Extended Data Fig. 9 | Hallux length and robusticity. a, Ratio (natural logarithm) of proximal hallucal phalanx total length to tibial physiologic length, relative to body mass (maximum femur head diameter). **b**, Box plots of hallux to femur head diameter ratios (natural logarithm). Box plots show the centre line (median), box limits (upper and lower quartiles), cross (arithmetic mean), whiskers (range) and individual values (circles). **c**, Size-adjusted hallucal phalanx midshaft robusticity (for explanation, see Extended Data Fig. 8c), relative to femur head diameter. All sample sizes (*n*) of biologically independent animals are reported in parentheses after the species names. For raw data, see Supplementary Tables 7, 22.



Extended Data Fig. 10 | *D. guggenmosi*, maxillary sinus and enamel thickness. a, Left maxilla with three-dimensional rendering of molar roots and maxillary sinus (blue) in lingual (left), anterior (middle) and occlusal (right) views. Sinus runs deep between the posterobuccal and lingual roots of M^2, rising anteriorly (dashed black line). Laterally the sinus extends deep into the zygomatic root (dashed white line). **b**, **c**, Enamel thickness measured on right M_2 (GPIT/MA 10000-03). Computed tomography image of the cross-section at distal sectional plane (**b**) and graphical conversion (**c**; grey, enamel; dark grey; dentine).

nature research

Corresponding author(s): Madelaine Böhme

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Data was collected on the original specimen using a phoenix v|tome|x s CT scanner at GeoZentrum Nordbayern (Friedrich-Alexander Universität Erlangen-Nürnberg, Germany) and a YXLON FF35 CT scanner at YXLON Inspection Service facility (Heidelberg / Germany). Comparative data of extant species were collected by using an Artec Space Spider surface scanner and Artec Studio software (versions 11-14).
Data analysis	For the micro CT-scan data analysis, we used Avizo 9.0 (ThermoFisher Scientific) and Geomagic Wrap 2017 (3D Systems Software) for the virtual reconstruction of longbones. 3D-prints were generated with Z-Suite 2.11 and printed on a Zortrax M200 FDM printer. Lateral tibia ellipse estimates we obtained using the Mathlab function “EllipseDirectFit” of Nikolai Chernov available from mathworks.com

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data generated or analysed during this study are included in the published article (and its supplementary information files). The CT-scans analysed during the current study are available from the corresponding author on reasonable request.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Morphologic description and functional interpretation of fossil hominid specimens.
Research sample	The research sample consists of 36 original fossil hominid bones/teeth from Hammerschmiede. The extant primates samples for skeletal comparison consists of about 350 adult and non-captive individuals of cercopithecids and hominids of both sexes.
Sampling strategy	No sample size calculation was performed. The sample size of fossils is limited by availability. The size of extant comparative samples (primates) varies between 10 and 60 individuals, which is a normal size in primatological anatomic comparisons.
Data collection	Data from the original fossil specimens were collected by M.B, D.R.B., N.S. and J.F. Micro-CT and surface scan data processing and collection was conducted by J.F. and A.T., in collaboration with A.S.D., U.K., T.L. and J.P.
Timing and spatial scale	Data collection started in spring 2018, followed by comparative data collection from summer 2018 to summer 2019.
Data exclusions	No data was excluded from the analysis.
Reproducibility	not applicable
Randomization	not applicable
Blinding	not applicable

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	not applicable for palaeontological excavations
Location	Hammerschmiede, Allgäu, Bavaria, southern Germany;), coordinates N 47° 55’ 38.5’’, E 10° 35.5’; fluvial channel of level HAM 5 at stratigraphic meter 12 in the local section, 685 m above sea level
Access and import/export	According to German (Bavarian) law no permissions needed for palaeontological excavations. Permission from the land owner have been obtained.
Disturbance	No disturbance (active mining pit)

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☐	☒ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Palaeontology

Specimen provenance	See above: According to German (Bavarian) law no permissions needed for palaeontological excavations. Non-formal permission from the land owner have been obtained.
Specimen deposition	All Hammerschmiede fossils are stored in the paleontological collection of the University of Tübingen (acronym GPIT), a research infrastructure of the Senckenberg Institute for Human Evolution and Palaeoenvironment (SHEP) Tübingen.
Dating methods	No new dates are provided.

☒ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Article

High-resolution lineage tracking reveals travelling wave of adaptation in laboratory yeast

https://doi.org/10.1038/s41586-019-1749-3

Received: 19 September 2018

Accepted: 4 October 2019

Published online: 13 November 2019

Alex N. Nguyen Ba[1,9], Ivana Cvijović[1,2,3,4,9], José I. Rojas Echenique[1,9], Katherine R. Lawrence[1,5], Artur Rego-Costa[1], Xianan Liu[6,7], Sasha F. Levy[6,7] & Michael M. Desai[1,3,4,8]*

In rapidly adapting asexual populations, including many microbial pathogens and viruses, numerous mutant lineages often compete for dominance within the population[1–5]. These complex evolutionary dynamics determine the outcomes of adaptation, but have been difficult to observe directly. Previous studies have used whole-genome sequencing to follow molecular adaptation[6–10]; however, these methods have limited resolution in microbial populations. Here we introduce a renewable barcoding system to observe evolutionary dynamics at high resolution in laboratory budding yeast. We find nested patterns of interference and hitchhiking even at low frequencies. These events are driven by the continuous appearance of new mutations that modify the fates of existing lineages before they reach substantial frequencies. We observe how the distribution of fitness within the population changes over time, and find a travelling wave of adaptation that has been predicted by theory[11–17]. We show that clonal competition creates a dynamical 'rich-get-richer' effect: fitness advantages that are acquired early in evolution drive clonal expansions, which increase the chances of acquiring future mutations. However, less-fit lineages also routinely leapfrog over strains of higher fitness. Our results demonstrate that this combination of factors, which is not accounted for in existing models of evolutionary dynamics, is critical in determining the rate, predictability and molecular basis of adaptation.

Rapidly adapting populations have complex evolutionary dynamics. In these systems, adaptation is not limited by the supply of mutations[18]. Instead, numerous beneficial mutations arise simultaneously and drive competing clonal expansions[7–9], often accompanied by neutral and deleterious hitchhiker mutations[6,19]. Studies have shown that this is the dominant mode of adaptation in many bacterial and viral pathogens[20–22], as well as in the somatic evolution of cancer[23] and immune repertoires[24]. In these contexts, clonal interference and hitchhiking have important consequences for the pace, outcomes and repeatability of evolution.

This mode of rapid adaptation cannot be described by classical evolutionary theory, because the fate of each mutation cannot be modelled in isolation[11,25,26]. Instead, selection acts on physically linked combinations of alleles, which leads to complex co-dependency between the fates of different mutations. This limits the efficiency of selection and renders evolution less predictable: strongly beneficial mutations can be outcompeted, whereas deleterious mutations in good genetic backgrounds can spread through the population[6,13,27].

Numerous studies have used whole-genome sequencing to investigate these effects in laboratory microbial populations[6–10], and have shown that clonal interference and hitchhiking are widespread. However, limitations on sequencing depth make it impractical to achieve a frequency resolution of higher than a few per cent, and barcoding-based methods[5,28,29] that offer better resolution are limited to short timescales. These limitations are critical in large microbial populations, in which theory suggests that the fates of mutations are often determined over long timescales by competition and hitchhiking among rare high-fitness lineages, and that the vast majority of driver mutations never reach detectable frequencies[11–17].

A renewable barcoding system

Here, we develop a renewable barcoding approach to observe evolutionary dynamics at high resolution on long timescales, by periodically adding new barcodes to split each tracked lineage into labelled 'sublineages' (Fig. 1a). Our approach uses three orthogonal *lox* sites: Cre-mediated recombination occurs between sites of the same type,

[1]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA. [2]Graduate Program in Systems Biology, Harvard University, Cambridge, MA, USA. [3]NSF-Simons Center for Mathematical and Statistical Analysis of Biology, Harvard University, Cambridge, MA, USA. [4]Quantitative Biology Initiative, Harvard University, Cambridge, MA, USA. [5]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, USA. [6]Joint Initiative for Metrology in Biology, SLAC National Accelerator Laboratory, Stanford University, Stanford, CA, USA. [7]Laufer Center for Physical and Quantitative Biology, Department of Biochemistry, Stony Brook University, Stony Brook, NY, USA. [8]Department of Physics, Harvard University, Cambridge, MA, USA. [9]These authors contributed equally: Alex N. Nguyen Ba, Ivana Cvijović, José I. Rojas Echenique. *e-mail: mdesai@oeb.harvard.edu

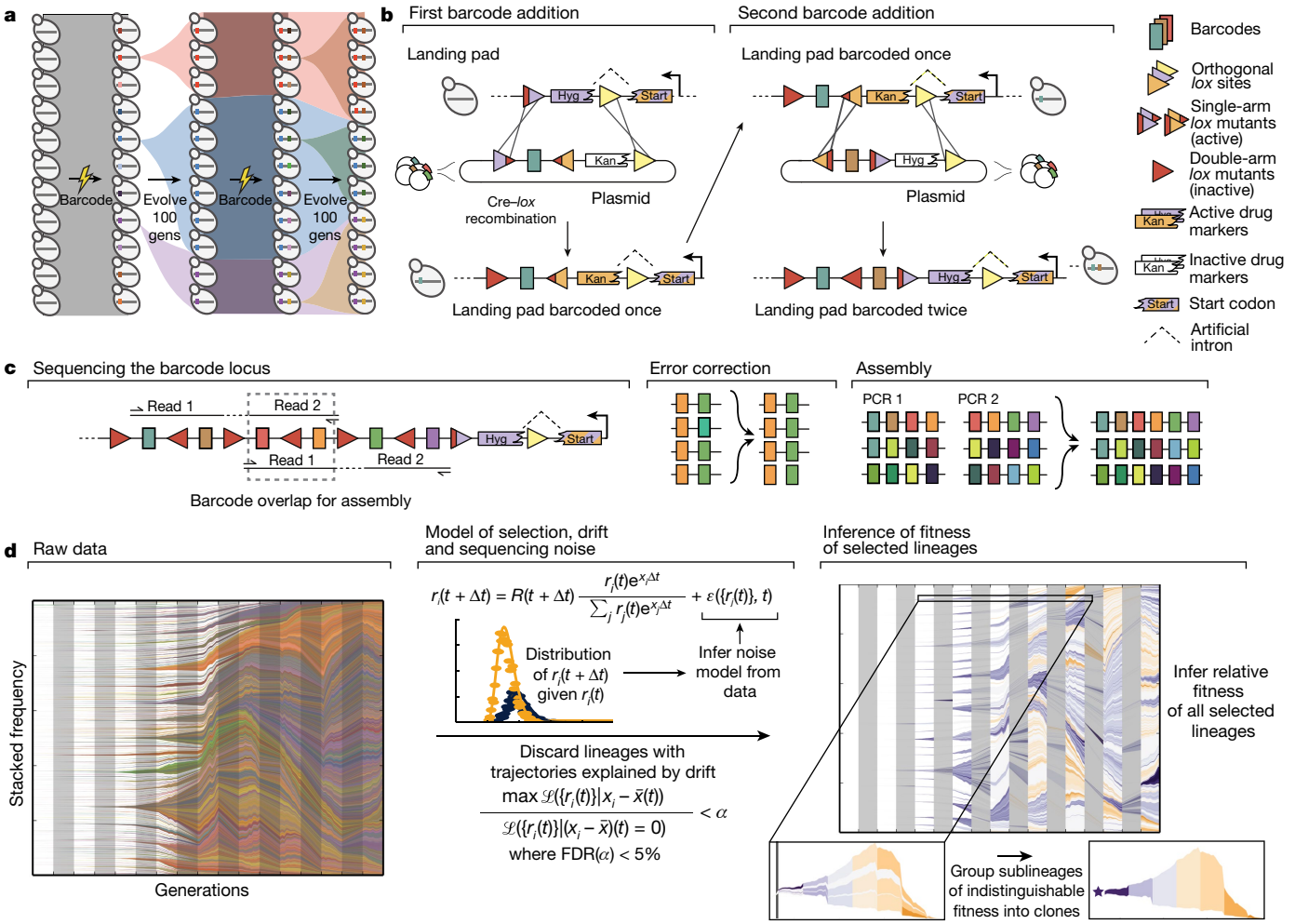


Fig. 1 | Renewable barcoding system and lineage dynamics. a, Experimental design. Diverse DNA barcodes are introduced into an initially clonal population, with each barcode labelling a small lineage. Every 100 generations (gens), we introduce new diverse barcodes immediately adjacent to existing barcodes, thereby subdividing each lineage into sublineages. **b**, Renewable barcoding system. The Cre–*lox* system consists of three orthogonal *lox* sites (coloured triangles), each of which can be modified with two arm disruptions (red) that are individually tolerated but jointly inactivating (Supplementary Information section 1). At each barcode addition, we combine arm disruptions to inactivate the old *lox* site, while adding a new orthogonal active *lox* site. Alternating *lox* orientations further limit undesired recombination. Drug markers (Kan, kanamycin (G418) resistance; Hyg, hygromycin B resistance) contain an intron 3′ splice-accepting site and must correctly integrate at the landing pad that contains the 5′ splice donor to be functional. **c**, When the barcode locus exceeds the length of an Illumina read, we use custom priming sites to sequence overlapping sets of four consecutive barcodes. After exploiting barcode diversity to identify and correct sequencing errors, we use these overlaps to unambiguously reconstruct the full barcode locus (Supplementary Information section 2). **d**, Inference pipeline. Left, raw barcode frequencies over time (left to right; colours chosen at random). For legibility, we only show lineages or sublineages with a frequency that exceeds 0.1% in at least one time point. Combined frequencies of lineages that do not individually reach 0.1% are shown as white space (or the colour of the parent when that parent exceeds a frequency of 0.1%). Middle, summary of the model used for identifying selected lineages (see Supplementary Information section 4 for details. In brief, we use the data to construct a parametric model for the strength of noise from genetic drift and sequencing and discard trajectories that are explained by noise alone, at a false discovery rate (FDR) of 5%. We then jointly infer the fitness of all remaining lineages and group lineages of indistinguishable fitness into clones (right).

but not between orthogonal types. Each site can be inactivated by two specific arm disruptions (one in each of the two Cre-binding regions), but retains high activity with only one disruption. We used this system to design barcoded plasmid libraries with complementary Cre–*lox* architecture, which we use to integrate barcodes at a designated genomic 'landing pad' locus (Fig. 1b, Supplementary Information section 1). At each barcode addition, Cre-mediated recombination combines arm disruptions to inactivate an old *lox* site, and adds a new orthogonal *lox* site with a single arm disruption to be used for the next barcode addition with a complementary plasmid library (Supplementary Information section 1). Each plasmid also contains an inactive drug marker that lacks a start codon; correct integration activates this marker by combining it with a start codon in the landing pad, separated by an artificial intron.

This system integrates new DNA barcodes immediately downstream of existing barcodes. Each individual thus acquires a string of barcodes that encode its ancestry, which can be read by sequencing. We read four barcodes per 150-bp paired-end Illumina read; when the barcode locus exceeds this length, we exploit overlapping fragments to assemble the complete locus (after using high barcode diversity to correct sequencing errors) (Fig. 1c, Supplementary Information sections 1.5, 2). This allows us to track the frequencies of all lineages and sublineages and hence trace the ancestry of the entire population.

Lineage tracking in evolving populations

We used this system to evolve two diploid yeast populations founded from identical clonal ancestors, each labelled with about 50,000 diverse

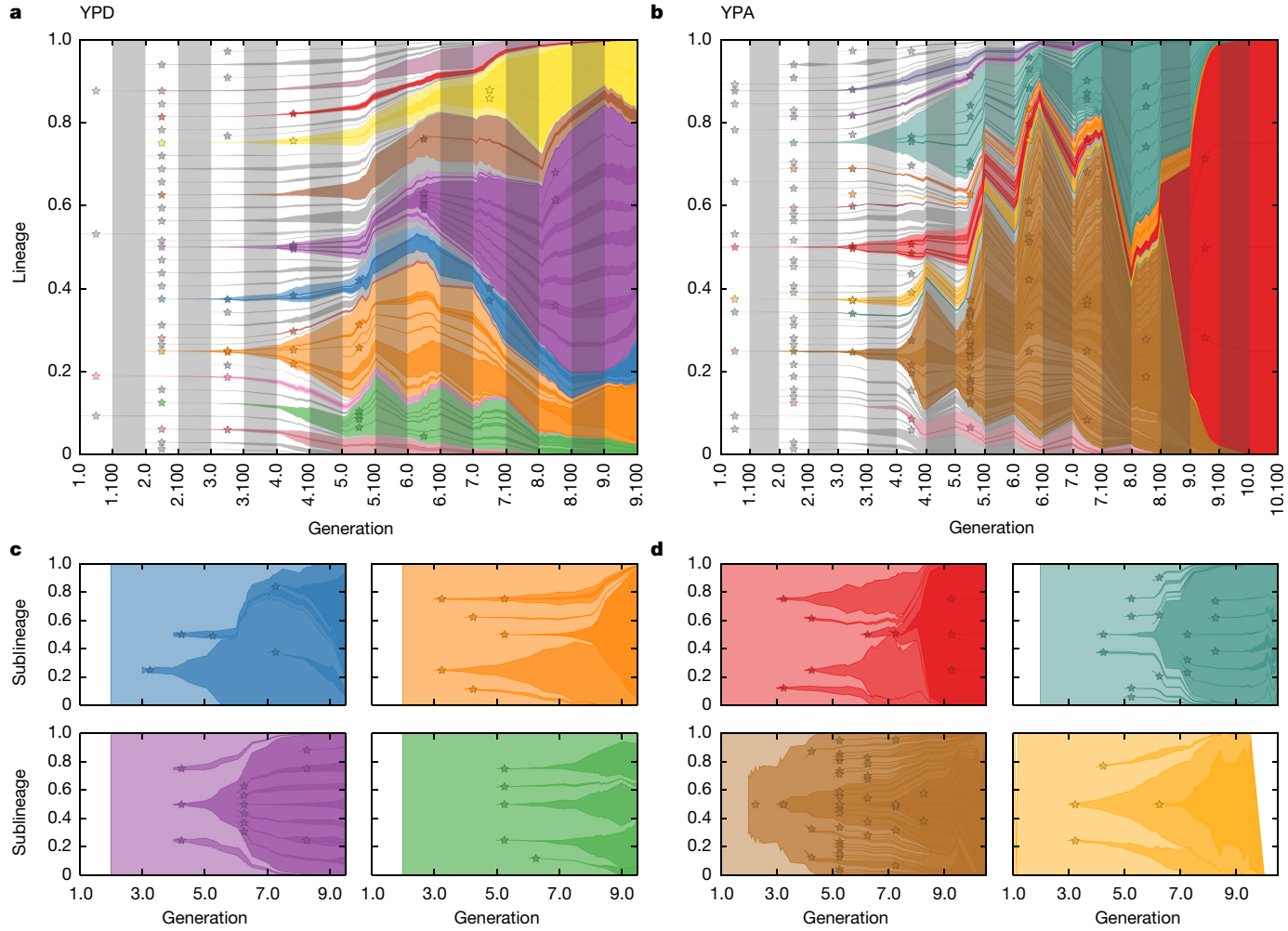


Fig. 2 | Inferred clonal dynamics. a, **b**, Muller diagrams showing dynamics of inferred beneficial mutations in YPD (**a**) and YPA (**b**) populations. Time is expressed in terms of epoch and generation (for example, 4.100 refers to generation 100 of epoch 4). Stars denote the establishment epoch of each new beneficial mutation (see Supplementary Information section 5). The opacity of the colours denotes the fitness of the corresponding lineage; mutant lineages that did not acquire additional beneficial mutations are grey. Grey bars denote barcoding intervals. **c**, **d**, Muller diagrams showing within-lineage dynamics in select lineages in the YPD (**c**) and YPA (**d**) populations. Colours are consistent with corresponding lineages in **a** and **b**. White space indicates periods during which the selected lineage was not observed.

barcodes. We maintained both populations in batch culture, with a 1:1,024 dilution every 24 h (10 generations per day with a bottleneck of about 500,000 cells; an effective size (N_e) of 5×10^6). An aliquot was frozen daily for analysis (Supplementary Information section 1.4). One population was maintained in rich medium (YPD) and the other in rich medium with 0.3% acetic acid (YPA), which leads to intracellular acidification that pilot studies have suggested leads to stronger selection pressures[30]. In studying these populations, our goal was to identify generic features of the evolutionary dynamics rather than details of differences between conditions. Our choice of environments maintains consistency with previous work, which indicates that these environments lead to rapid adaptation involving rich dynamics that could not be observed using earlier approaches[5–8].

We re-barcoded each population every 100 generations with about 50,000 additional unique barcodes. This diversity was chosen to ensure that barcoding does not introduce a substantial bottleneck; at 10% of the daily bottleneck every 10 days, it does not change the scale of genetic drift or the effective population size (Supplementary Information section 4.4). It also ensures that we can detect relevant selection pressures that act on lineages once those lineages become large enough that their dynamics are not dominated by drift (Supplementary Information section 4.4). However, we note that although our barcoding procedure is designed to be minimally perturbative, it does involve propagation and selection steps. Thus, strictly speaking we are studying evolution in a fluctuating environment that alternates between 'evolution' and 'barcoding' conditions—although, as we see below, the role of these fluctuations is minor.

After 1,000 generations of evolution (ten 100-generation 'epochs'), we sequenced the barcode locus at a depth of around 10^5 reads in every frozen time point. This yielded 110 sequenced time points per population (11 time points per epoch at 10-generation intervals, although we excluded the final epoch of the YPD population owing to barcoding failure; see Supplementary Information section 1.4). We use this data to infer which lineages contain mutations that are beneficial in either evolution or barcoding conditions, and we exploit phylogeny to infer in which epoch each mutation was established (that is, within 100-generation resolution; Fig. 1d, Supplementary Information section 5). This allows us to group barcodes into 'clones', each founded by a new mutation. We then jointly infer the fitness effects of all mutations in evolution and barcoding conditions. Because we barcode frequently, the dynamics are determined by the average fitness across the two conditions (Supplementary Information section 6.2). We therefore use this average fitness for the analysis below (although simply neglecting

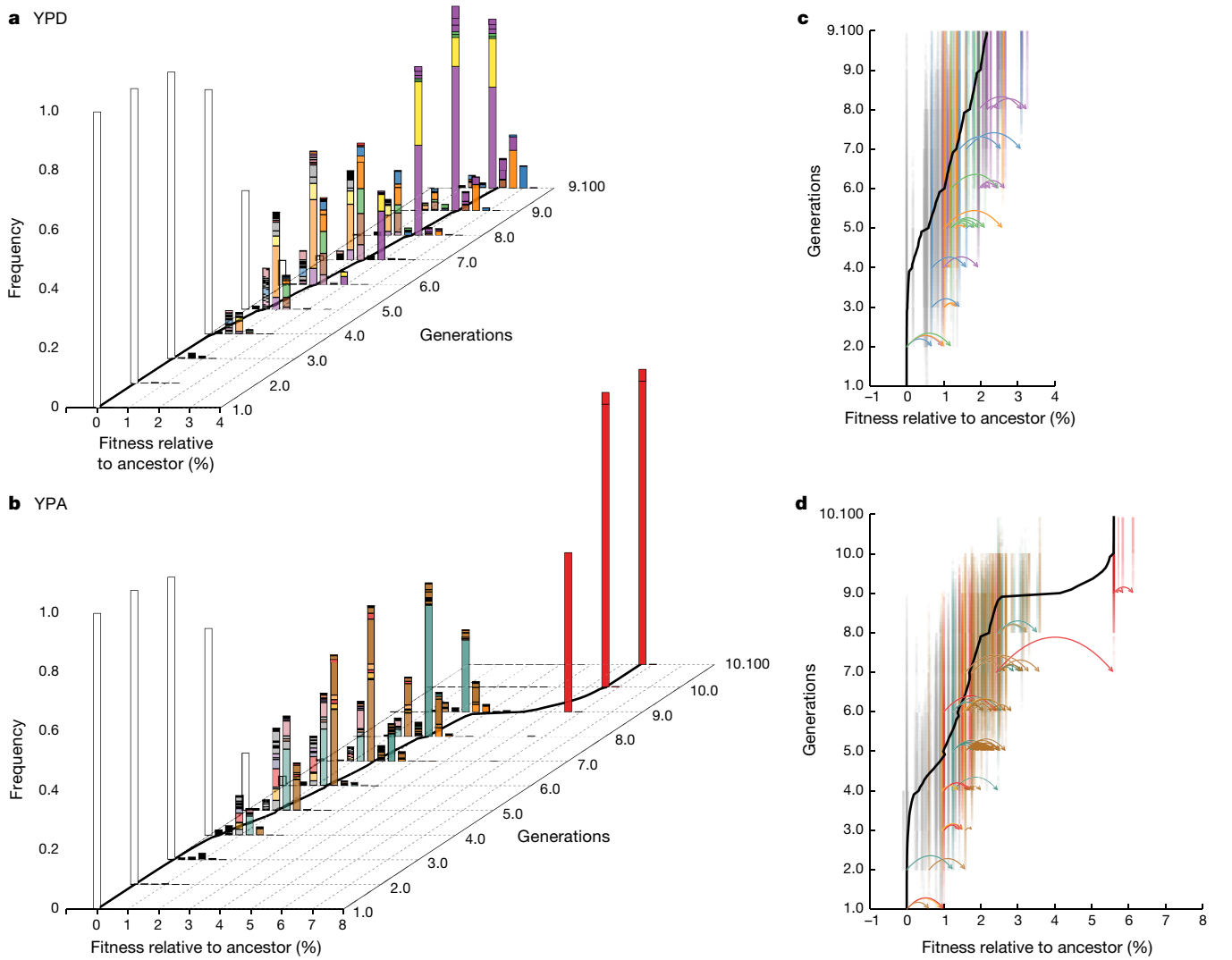


Fig. 3 | Travelling wave dynamics. a, **b**, Inferred distribution of fitness within the YPD (**a**) and YPA (**b**) populations over time. All fitness are the average across evolution and barcoding conditions (Supplementary Information 6.2). Each coloured bar denotes the frequency and fitness of a corresponding lineage in Fig. 2. White bars correspond to the ancestor. Black lines denote the inferred mean fitness of the population. **c**, **d**, Genealogical relationships among lineages in the YPD (**c**) and YPA (**d**) populations show frequent leapfrogging events. Each clonal lineage is shown at its corresponding fitness. The opacity of the colours indicates the frequency of the lineage. Colours of the lineages shown in Fig. 2c, d are consistent with that figure; all other lineages are grey. Mutational events within highlighted lineages are shown as arrows; each event arises in one clonal lineage and founds a new lineage at a new fitness.

the barcoding environment leads to qualitatively similar conclusions; see Supplementary Information section 6).

Our ability to detect mutations is limited primarily by genetic drift. We cannot identify mutations until they are common enough that their fitness effects lead to frequency changes larger than this stochastic force (which typically corresponds to lineages at frequencies greater than 10^{-4}). Because fitness inference requires sufficient time-course data, we are also unable to detect most mutations that arise in the final 100–200 generations of the experiment (Supplementary Information section 5.3). Our analysis thus only identifies a subset of beneficial mutations, and our clones are clonal only with respect to these.

We find that in both populations, many beneficial mutations arise early in the experiment, founding clones that compete for dominance (Fig. 2a, b). Some of these clones diversify through further beneficial mutations, and a handful obtain multiple mutations, which interfere with one another within the parent clone (Fig. 2c, d). In some cases we observe multiple nested interference events (Fig. 2c, d). All but the largest of these events are undetectable by metagenomic sequencing at approximately 25× depth (which corresponds to about the same total number of sequencing reads as our barcode data; Extended Data Fig. 1a, b).

We can also visualize how the fitness composition of the population changes over time (Fig. 3). The population initially diversifies as numerous beneficial mutations arise on the ancestral background, creating a distribution of fitness within the population (Fig. 3a, b). As these clones expand, the mean fitness of the population increases (Fig. 3, Extended Data Fig. 2), causing less-fit lineages to fall behind and begin to die out. However, diversity is maintained by new beneficial mutations, which continuously create clones of even higher fitness (Fig. 3c, d). This maintenance of diversity in the face of strong selection is an expected feature of rapid adaptation that has been predicted by theory[11–17] but not previously observed directly.

Determinants of lineage success

These dynamics lead to a complex picture of the determinants of success of individual lineages. In the absence of further mutations, the

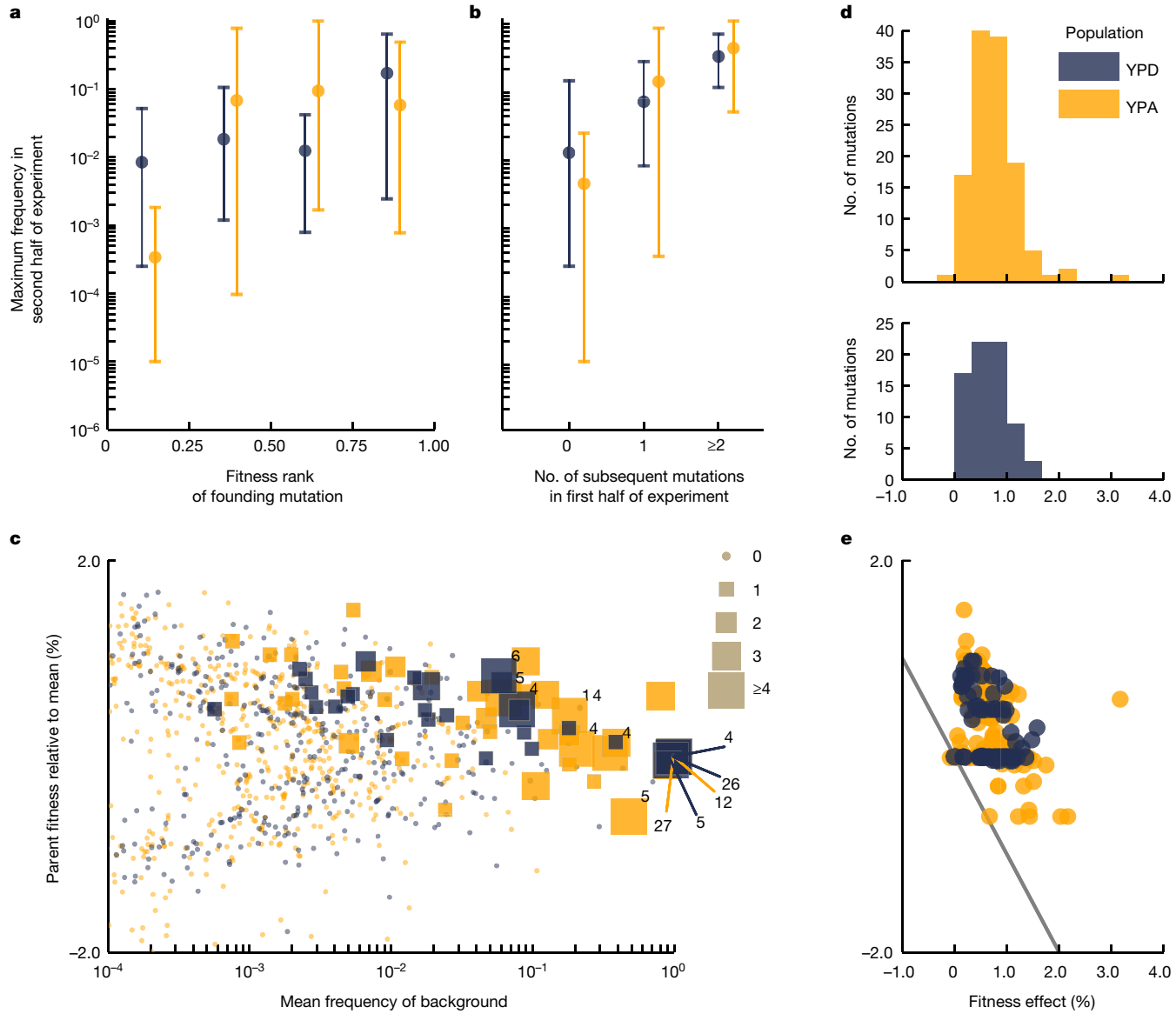


Fig. 4 | Travelling wave dynamics and factors determining the success of mutant lineages. a, Relationship between initial within-population fitness rank of a mutation that arises in the ancestral background and its maximum frequency in the second half of the experiment (using the second half avoids confounding axes in **b**). $n = 35$ and $n = 47$ unique lineages in YPD and YPA respectively. Dots represent the mean, and lines show the range of maximum frequencies in each founding fitness quantile. **b**, Relationship between the number of subsequent beneficial mutations landing on the founding clonal background of a lineage (in the first half of the experiment) and the eventual maximum frequency of that lineage (in the second half of the experiment). **c**, Effect of lineage frequency and fitness on the likelihood of acquiring additional beneficial mutations. Each point represents the mean frequency and fitness of a lineage in a given 100-generation epoch; symbol size denotes how many additional beneficial mutations that lineage acquired (numbers indicate lineages that acquire more than four). **d**, Histograms of effect sizes of all inferred mutations. **e**, Effect sizes of mutations arising on parental backgrounds as a function of mean parental relative fitness in the epoch in which each mutation arose. The region below the grey line corresponds to mutations that would create lineages less fit than the current mean fitness.

fitness of a lineage should be the only predictor of its success. Yet we find that the initial fitness of a mutant lineage is only a modest predictor of its fate (Fig. 4a). Another key factor is whether a lineage acquires further beneficial mutations (Fig. 4b). Although this is influenced by fitness (see below), even high-fitness lineages that do not acquire further beneficial mutations are readily outcompeted, and lower-fitness lineages that acquire multiple mutations can succeed (Extended Data Fig. 3). The likelihood of a lineage acquiring further beneficial mutations is in turn affected by two main factors (Fig. 4c). First, larger lineages have more opportunities to acquire beneficial mutations. Second, the fitness of a lineage has a critical role: mutations that arise in a highly fit and hence rapidly expanding lineage will be less likely to be lost to genetic drift. Thus, highly fit backgrounds can accumulate beneficial mutations of both strong and weak effect, whereas only rare strong mutations can establish on lower-fitness backgrounds. Consistent with this, our data show that high-fitness backgrounds acquire both weakly and strongly beneficial mutations, but low-fitness backgrounds only acquire the latter (Fig. 4d, e). This means that fitter backgrounds have access to a larger number of beneficial mutations, creating a rich-get-richer effect that can lead to bursts of mutations at the expanding front of the fitness wave. These bursts arise owing to dynamical considerations, and are not in themselves evidence of historical contingency as a result of mutator phenotypes or other modifiers of adaptability (Supplementary Information section 6.5).

These results are qualitatively consistent with recent theory that suggests that rapidly evolving populations can be described by 'travelling wave' models[11–17]. In this picture, mutations continuously generate variation in fitness while selection destroys it by eliminating less-fit genotypes, leading to a broad distribution of fitness around an increasing mean (a fitness wave). However, these models have only been analysed in parameter regimes in which the future common ancestor of a population is always one of the fittest lineages (although see one previous study that discusses scenarios in which this can be violated[17]). Instead, clonal competition in our experiment is characterized by routine 'leapfrogging' events, in which lineages of initially low relative fitness acquire strong beneficial mutations that pull them to prominence, causing dramatic reversals of fate. For example, in the YPD population (Fig. 3c) the green lineage—which is the fittest at the start—is leapfrogged by the orange, blue and purple backgrounds; the blue lineage then falls behind, only to later leapfrog all others. Similarly, in the YPA population (Fig. 3d), the brown lineage appears to outcompete the turquoise, red and yellow lineages, only to be leapfrogged by two strongly beneficial mutations in a red lineage that is initially much less fit (replay experiments validate this event; see Supplementary Information section 6.3).

Leapfrogging events not only alter the fates of individual lineages, but also cause fluctuations in the fitness distribution and modulate the pace of adaptation. Both within-population fitness and genetic variation increase during initial diversification before reaching a plateau as a travelling wave is established (Fig. 3, Extended Data Fig. 4). However, leapfrogging can cause fluctuations in this travelling wave: the creation of a lineage with anomalously high fitness can lead to a reduction in diversity at first, but at the same time enable rapid further diversification within this lineage that later re-establishes variation (Fig. 3c, d, Extended Data Fig. 4). These fluctuations affect the success of any individual mutation and the dynamics of the travelling wave, and hence have a major role in determining the outcomes of evolution.

Discussion

Previous theory has assumed that the effects of leapfrogging and fluctuations are occasional perturbations that can be largely ignored[11–17]. Our results suggest that they instead have a central role. Although our system involves microbial populations of modest size, the importance of these effects is expected to depend only weakly on population size and mutation rate (because relevant timescales only depend logarithmically on these quantities[12,17]). Thus our results suggest that leapfrogging and fluctuations may be routine in the evolution of a wide range of microorganisms and viruses. A new theoretical framework is essential to develop accurate models of evolution in these systems. The renewable barcoding approach we have introduced here offers the potential to test these models, and to observe evolutionary dynamics in a variety of contexts at sufficient resolution to investigate the role of other factors such as frequency-dependent selection or mutations that alter the adaptability of individual lineages.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Raw sequencing reads have been deposited in the NCBI BioProject database under accession number PRJNA559526. All associated metadata, as well as the source code for the sequencing pipeline, downstream analyses, and figure generation, are available at GitHub (https://github.com/icvijovic/lineage-tracking). Source Data for Figs. 2–4 are provided with the paper.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1749-3.

1. Gerrish, P. J. & Lenski, R. E. The fate of competing beneficial mutations in an asexual population. *Genetica* **102–103**, 127–144 (1998).
2. Desai, M. M., Fisher, D. S. & Murray, A. W. The speed of evolution and maintenance of variation in asexual populations. *Curr. Biol.* **17**, 385–394 (2007).
3. Miller, C. R., Joyce, P. & Wichman, H. A. Mutational effects and population dynamics during viral adaptation challenge current models. *Genetics* **187**, 185–202 (2011).
4. De Visser, J. A. G. M. et al. Diminishing returns from mutation supply rate in asexual populations. *Science* **283**, 404–406 (1999).
5. Levy, S. F. et al. Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **519**, 181–186 (2015).
6. McDonald, M. J., Rice, D. P. & Desai, M. M. Sex speeds adaptation by altering the dynamics of molecular evolution. *Nature* **531**, 233–236 (2016).
7. Lang, G. I. et al. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* **500**, 571–574 (2013).
8. Kvitek, D. J. & Sherlock, G. Whole genome, whole population sequencing reveals that loss of signaling networks is the major adaptive strategy in a constant environment. *PLoS Genet.* **9**, e1003972 (2013).
9. Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E. & Desai, M. M. The dynamics of molecular evolution over 60,000 generations. *Nature* **551**, 45–50 (2017).
10. Tenaillon, O. et al. Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature* **536**, 165–170 (2016).
11. Neher, R. A. Genetic draft, selective interference, and population genetics of rapid adaptation. *Annu. Rev. Ecol. Evol. Syst.* **44**, 195–215 (2013).
12. Desai, M. M. & Fisher, D. S. Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics* **176**, 1759–1798 (2007).
13. Good, B. H., Rouzine, I. M., Balick, D. J., Hallatschek, O. & Desai, M. M. Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. *Proc. Natl Acad. Sci. USA* **109**, 4950–4955 (2012).
14. Rouzine, I. M., Brunet, E. & Wilke, C. O. The traveling-wave approach to asexual evolution: Muller's ratchet and speed of adaptation. *Theor. Popul. Biol.* **73**, 24–46 (2008).
15. Tsimring, L. S., Levine, H. & Kessler, D. A. RNA virus evolution via a fitness-space model. *Phys. Rev. Lett.* **76**, 4440–4443 (1996).
16. Hallatschek, O. The noisy edge of traveling waves. *Proc. Natl Acad. Sci. USA* **108**, 1783–1787 (2011).
17. Fisher, D. S. Asexual evolution waves: fluctuations and universality. *J. Stat. Mech.* **2013**, P01011 (2013).
18. Cvijović, I., Nguyen Ba, A. N. & Desai, M. M. Experimental studies of evolutionary dynamics in microbes. *Trends Genet.* **34**, 693–703 (2018).
19. Buskirk, S. W., Peace, R. E. & Lang, G. I. Hitchhiking and epistasis give rise to cohort dynamics in adapting populations. *Proc. Natl Acad. Sci. USA* **114**, 8330–8335 (2017).
20. Zanini, F. et al. Population genomics of intrapatient HIV-1 evolution. *eLife* **4**, e11282 (2015).
21. Lieberman, T. D. et al. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat. Genet.* **43**, 1275–1280 (2011).
22. Strelkowa, N. & Lässig, M. Clonal interference in the evolution of influenza. *Genetics* **192**, 671–682 (2012).
23. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
24. Nourmohammad, A., Otwinowski, J., Luksza, M., Mora, T. & Walczak, A. M. Fierce selection and interference in B-cell repertoire response to chronic HIV-1. *Mol. Biol. Evol.* **36**, 2184–2194 (2019).
25. Muller, H. Some genetic aspects of sex. *Am. Nat.* **66**, 118–138 (1932).
26. Maynard Smith, J. Evolution in sexual and asexual populations. *Am. Nat.* **102**, 469–473 (1968).
27. Good, B. H. & Desai, M. M. Deleterious passengers in adapting populations. *Genetics* **198**, 1183–1208 (2014).
28. Blundell, J. R. & Levy, S. F. Beyond genome sequencing: lineage tracking with barcodes to study the dynamics of evolution, infection, and cancer. *Genomics* **104**, 417–430 (2014).
29. Blundell, J. R. et al. The dynamics of adaptive genetic diversity during the early stages of clonal evolution. *Nat. Ecol. Evol.* **3**, 293–301 (2019).
30. Ludovico, P., Sousa, M. J., Silva, M. T., Leão, C. & Côrte-Real, M. *Saccharomyces cerevisiae* commits to a programmed cell death process in response to acetic acid. *Microbiology* **147**, 2409–2415 (2001).

Acknowledgements We thank E. Jerison, A. Moses, A. Murray and members of the M.M.D. laboratory for comments on the manuscript. A.N.N.B. acknowledges support from NSERC; I.C. acknowledges support from the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard University (NSF grant DMS-1764269) and the Harvard FAS Quantitative Biology Initiative; K.R.L. acknowledges support from the Fannie and John Hertz Foundation Graduate Fellowship Award and the NSF Graduate Research Fellowship Program; S.F.L. acknowledges support from the NIH (grants HG008354 and HL127522); M.M.D. acknowledges support from the Simons Foundation (grant 376196), the NSF (grant DEB-1655960) and the NIH (grant GM104239). Computational work was performed on the Odyssey cluster supported by the Research Computing Group at Harvard University.

Author contributions A.N.N.B., J.I.R.E. and M.M.D. designed the project; A.N.N.B. and J.I.R.E. constructed the barcoding system with assistance from X.L., S.F.L. and M.M.D.; A.N.N.B., J.I.R.E., K.R.L. and A.R.-C. conducted the experiments; A.N.N.B., J.I.R.E. and I.C. designed and conducted the bioinformatics analysis; I.C. developed theory and inference methods and analysed the data; I.C., M.M.D., A.N.N.B. and J.I.R.E. wrote the paper.

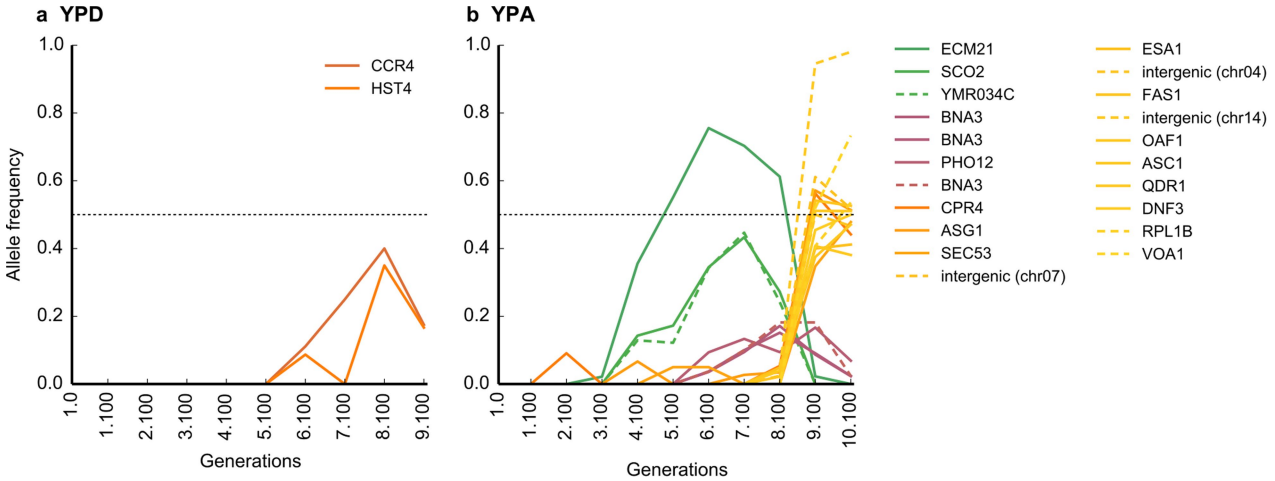
Competing interests The authors declare no competing interests.

Additional information

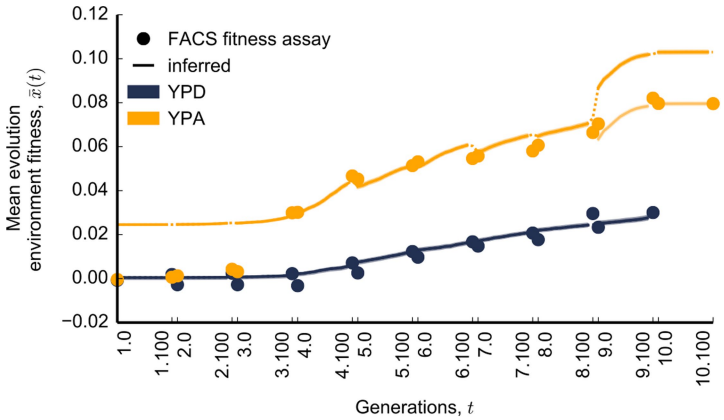
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1749-3.

Correspondence and requests for materials should be addressed to M.M.D.

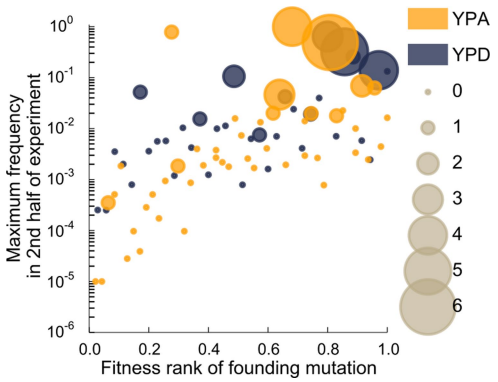
Peer review information *Nature* thanks David Gresham, Daniel Weinreich and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



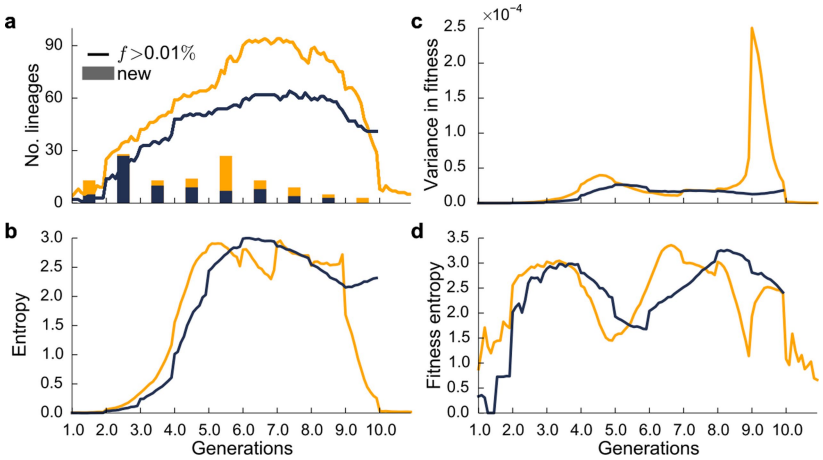
Extended Data Fig. 1 | Allele frequency trajectories in the two populations, as detected by metagenomic sequencing. a, **b**, In both the YPD (**a**) and the YPA (**b**) population, solid lines denote missense and nonsense mutations, and dotted lines denote synonymous mutations and those falling in intergenic regions. Lines are coloured according to the peak time of the trajectory. Note that a frequency of 50% (dotted line) corresponds to a mutation that fixes as a heterozygote.



Extended Data Fig. 2 | Comparison of inferred and measured population mean fitness trajectories. All fitness measurements and inferences refer to the evolution environment only. Trajectories have been offset to agree with the fitness assay at time point 3.100. Dots denote barcoding intervals. Shaded regions around the trajectories denote estimates of 95% confidence intervals for the inferred mean fitness trajectory, which often do not exceed the width of the lines (Supplementary Information section 6.1). In the case of the YPA population, lighter colours denote mean fitness trajectories over the last two epochs, offset to agree with fitness assays in the last time point (see Supplementary Information section 6.6 for a discussion of potential reasons for these discrepancies) FACS, fluorescence-activated cell sorting.



Extended Data Fig. 3 | Predictors of the success of lineages. The size of each dot denotes the number of later beneficial mutations that occur in the founding clonal background of a lineage (in the first half of the experiment).



Extended Data Fig. 4 | Genetic variation over time. a, Total number of lineages above a threshold frequency (0.01%) over time. Bars denote the number of new lineages that arise in each 100-generation interval. **b**, Genetic diversity within each population over time, as measured by entropy (Supplementary Information section 6.4). **c**, Variance in fitness over time. **d**, Fitness diversity within each population over time, as measured by fitness entropy. Fitness entropy quantifies how fitness variance is distributed among lineages (Supplementary Information section 6.4).

nature research

Corresponding author(s): Michael M Desai

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	Freely available third-party software trimmomatic v0.35 and breseq v0.27.1b as well as custom scripts were used for data analysis. All custom algorithms are described in the SI Appendix; the source code is available on GitHub (https://github. com/icvijovic/lineage-tracking).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw sequencing reads have been deposited in the NCBI BioProject database under accession number PRJNA559526. All associated metadata, as well as the source code for the sequencing pipeline, downstream analyses, and figure generation, are available at GitHub (https://github. com/icvijovic/lineage-tracking).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size of a single population per evolutionary condition is sufficient, because our method makes it possible to detect a large number of mutations in each condition, allowing for statistical comparisons between their evolutionary dynamics and fates.
Data exclusions	The last 100 generations of evolution in the YPD environment were excluded from further analysis, because of a failure of the re-barcoding procedure, as described in the SI Appendix. The exclusion criteria were not pre-established. No other data were excluded.
Replication	No additional replication was performed.
Randomization	The clonal ancestors for the two evolution experiments were selected randomly from a single isogenic population.
Blinding	Blinding to the experimental environment was not relevant to our study because the two populations were subjected to identical data collection and data analysis procedures.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☐	☒ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Unique biological materials

Policy information about availability of materials

Obtaining unique materials	All unique materials are readily available from the authors.

Article

Marine Proteobacteria metabolize glycolate via the β-hydroxyaspartate cycle

https://doi.org/10.1038/s41586-019-1748-4

Received: 29 March 2019

Accepted: 20 September 2019

Published online: 13 November 2019

Lennart Schada von Borzyskowski[1]*, Francesca Severi[1], Karen Krüger[2], Lucas Hermann[3], Alexandre Gilardet[1], Felix Sippel[1], Bianca Pommerenke[1], Peter Claus[1], Niña Socorro Cortina[1], Timo Glatter[4], Stefan Zauner[5], Jan Zarzycki[1], Bernhard M. Fuchs[2], Erhard Bremer[3,6], Uwe G. Maier[5,6], Rudolf I. Amann[2] & Tobias J. Erb[1,6]*

One of the most abundant sources of organic carbon in the ocean is glycolate, the secretion of which by marine phytoplankton results in an estimated annual flux of one petagram of glycolate in marine environments[1]. Although it is generally accepted that glycolate is oxidized to glyoxylate by marine bacteria[2–4], the further fate of this C_2 metabolite is not well understood. Here we show that ubiquitous marine Proteobacteria are able to assimilate glyoxylate via the β-hydroxyaspartate cycle (BHAC) that was originally proposed 56 years ago[5]. We elucidate the biochemistry of the BHAC and describe the structure of its key enzymes, including a previously unknown primary imine reductase. Overall, the BHAC enables the direct production of oxaloacetate from glyoxylate through only four enzymatic steps, representing—to our knowledge—the most efficient glyoxylate assimilation route described to date. Analysis of marine metagenomes shows that the BHAC is globally distributed and on average 20-fold more abundant than the glycerate pathway, the only other known pathway for net glyoxylate assimilation. In a field study of a phytoplankton bloom, we show that glycolate is present in high nanomolar concentrations and taken up by prokaryotes at rates that allow a full turnover of the glycolate pool within one week. During the bloom, genes that encode BHAC key enzymes are present in up to 1.5% of the bacterial community and actively transcribed, supporting the role of the BHAC in glycolate assimilation and suggesting a previously undescribed trophic interaction between autotrophic phytoplankton and heterotrophic bacterioplankton.

Global net primary production has been estimated to be approximately 100 petagrams of carbon per year, equal parts of which are produced in terrestrial and marine habitats[6]. In the oceans, more than a third of primary production can be released into the water column by phytoplankton as dissolved organic carbon[7], generating a plethora of substrates for heterotrophic bacterioplankton. An abundant component of the pool of dissolved organic carbon is the carboxylic acid glycolate, which is released as a photorespiratory waste product of marine autotrophs[3,8,9]. Concentrations of glycolate in the nanomolar-to-low micromolar range have been measured in different marine habitats[1,2,10,11] (Extended Data Fig. 1), and the compound is readily taken up by bacterioplankton[12]. The first step in glycolate metabolism is its oxidation to glyoxylate, which is catalysed by the enzyme glycolate oxidase. The abundance and transcription of the *glcD* gene, which encodes a subunit of glycolate oxidase, has previously been used to investigate bacterial groups that are capable of glycolate utilization[4,13]. However, it has been assumed that glycolate is the subject of bacterial oxidation mainly to conserve energy[2–4]; the further fate of glyoxylate has not been described in detail. For SAR11 bacteria, it has been shown that glyoxylate can be used to replace the obligate glycine requirement[14]. In SAR11 and other bacteria, glyoxylate can be co-assimilated by malate synthase into the tricarboxylic acid cycle[14–16] or directly assimilated into central carbon metabolism through the well-studied glycerate pathway[17,18]. An alternative solution is the BHAC[5,19], which has been previously proposed to operate in the Alphaproteobacterium *Paracoccus denitrificans*[20,21]. However, the complete reaction sequence and the proteins comprising this pathway and their detailed biochemistry have remained unknown for the past 56 years.

On the basis of the sequence of a putative β-hydroxyaspartate aldolase gene (*dhaa*; GenBank accession number AB075600) from *P. denitrificans* IFO 13301[22], we identified a homologue in the genome of *P. denitrificans* DSM413 (BLT64_RS06500), annotated as a DSD1 family pyridoxal 5-phosphate (PLP)-dependent enzyme. This gene is part of a gene cluster, which consists of four structural genes and a putative transcriptional regulator that we termed *bhcABCD* and *bhcR* (Fig. 1a). In addition to the gene for the putative β-hydroxyaspartate aldolase (*bhcC*), the cluster comprises the open reading frames that encode a putative PLP-dependent aminotransferase (BLT64_RS06510, *bhcA*), a putative serine/threonine dehydratase (BLT64_RS06505, *bhcB*) and a putative ornithine cyclodeaminase (BLT64_RS06495, *bhcD*). The

[1]Department of Biochemistry & Synthetic Metabolism, Max Planck Institute for Terrestrial Microbiology, Marburg, Germany. [2]Department of Microbial Ecology, Max Planck Institute for Marine Microbiology, Bremen, Germany. [3]Laboratory for Molecular Microbiology, Department of Biology, Philipps-University Marburg, Marburg, Germany. [4]Facility for Mass Spectrometry and Proteomics, Max Planck Institute for Terrestrial Microbiology, Marburg, Germany. [5]Laboratory for Cell Biology, Department of Biology, Philipps-University Marburg, Marburg, Germany. [6]LOEWE-Center for Synthetic Microbiology, Philipps-University Marburg, Marburg, Germany. *e-mail: schada@mpi-marburg.mpg.de; toerb@mpi-marburg.mpg.de

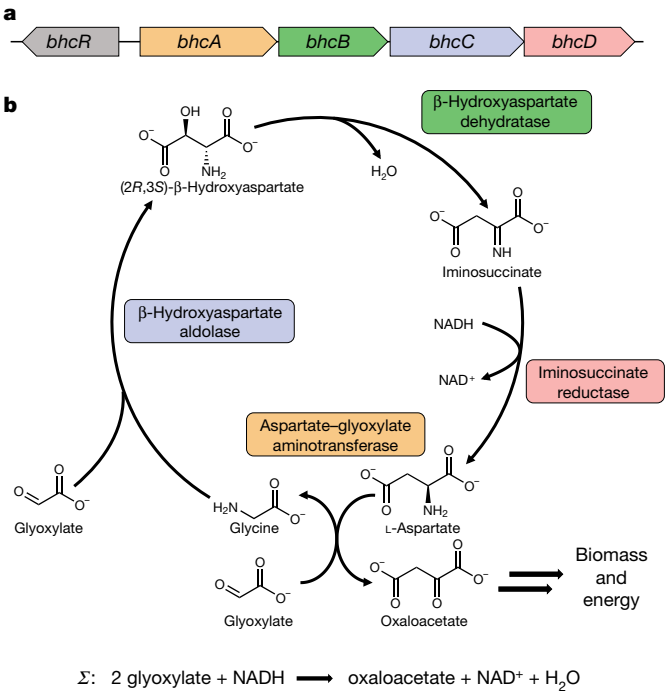


Fig. 1 | The BHAC. a, Genetic structure of the bhc gene cluster in *P. denitrificans* DSM 413. **b**, Reaction sequence and net balance of the BHAC. **c**, Cartoon representation of the β-hydroxyaspartate aldolase (BhcC) homodimer with superimposed protein surface (PDB 6QKB). **d**, Cartoon representation of the iminosuccinate reductase (BhcD) homodimer with superimposed protein surface (PDB 6RQA).

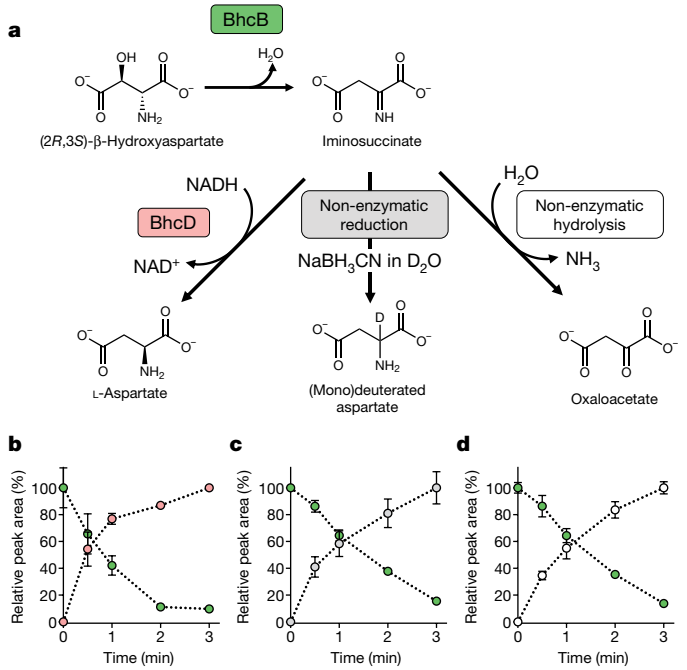


Fig. 2 | Reaction sequence catalysed by β-hydroxyaspartate dehydratase (BhcB) and iminosuccinate reductase (BhcD). a, Overview of the relevant reactions. **b**, Production of L-aspartate (red) from (2*R*, 3*S*)-β-hydroxyaspartate (green) by BhcB and reduction of iminosuccinate by BhcD. **c**, Production of (mono)deuterated aspartate (grey) from (2*R*, 3*S*)-β-hydroxyaspartate (green) by BhcB and reduction of iminosuccinate via $NaBH_3CN$ in D_2O. The data represent the formation of monodeuterated aspartate; owing to proton exchange, di- and trideuterated aspartate can also be formed in small quantities. **d**, Production of oxaloacetate (white) from (2*R*, 3*S*)-β-hydroxyaspartate (green) by BhcB and subsequent hydrolysis of iminosuccinate when neither BhcD nor $NaBH_3CN$ are added. **b**–**d**, Data are mean ± s.d.; $n = 3$ independent experiments.

putative transcriptional regulator (BLT64_RS06515), annotated as IclR-family regulator, is located in the opposite orientation to the four structural genes.

We expressed and characterized the four enzymes that are encoded in the gene cluster. BhcA is a PLP-dependent aminotransferase that transaminates glyoxylate into glycine using aspartate as the preferred amino group donor. BhcB functions as a β-hydroxyaspartate dehydratase. BhcC is a β-hydroxyaspartate aldolase, the key enzyme of the BHAC that catalyses the condensation of glyoxylate and glycine into β-hydroxyaspartate. This enzyme is closely related to D-threonine aldolases (Extended Data Fig. 2). The crystal structure of β-hydroxyaspartate aldolase that we solved at 1.7 Å (Protein Data Bank (PDB) 6QKB) shows that the three amino acids A160, A195 and S313 distinguish the active site of BhcC from that of D-threonine aldolases, providing a signature sequence for this enzyme family (Fig. 1c, Extended Data Fig. 2 and Extended Data Table 1). When combined, the BhcABC proteins were sufficient to reconstruct a reaction sequence from aspartate and two molecules of glyoxylate to two molecules of oxaloacetate and free ammonia. However, this left us puzzled about the function of the fourth open reading frame, the putative ornithine cyclodeaminase (*bhcD*).

When we tested BhcD in combination with BhcB, we discovered that it functions as an imine reductase (IRED) that accepts a labile iminosuccinate intermediate[23] formed by the latter enzyme (Fig. 2a, b). We used sodium cyanoborohydride trapping to demonstrate that BhcB produces iminosuccinate (Fig. 2c). Although this compound spontaneously decays into free ammonia and oxaloacetate in solution (Fig. 2d), iminosuccinate is reduced to L-aspartate in the presence of BhcD, thereby regenerating the amino group donor for the first step of the BHAC. IREDs are extensively investigated owing to their biotechnological potential[24]. Almost all IREDs described to date act on secondary imines, whereas the reduction of a free primary imine—as catalysed by BhcD—has not previously been described. The enzymatic reduction of primary imines is known only as part of the reaction sequence in glutamate dehydrogenase[25] and as part of a non-physiological side reaction of ketimine reductases[26]. The crystal structure of BhcD, which we solved to a resolution of 2.6 Å (PDB 6RQA), shows major differences in the active site compared to L-alanine dehydrogenase from *Archaeoglobus fulgidus* (PDB 1OMO), the closest structural homologue within the ornithine cyclodeaminase/μ-crystalline enzyme superfamily (Fig. 1d, Extended Data Fig. 3 and Extended Data Table 1). Phylogenetic analysis supports these active site differences and reveals that BhcD and its homologues constitute a novel family of primary IREDs within the ornithine cyclodeaminase/μ-crystalline superfamily (Extended Data Fig. 3).

Table 1 | Kinetic parameters of the four enzymes of the BHAC

Enzyme	Substrate	k_{cat} (s^{-1})	App. K_M (mM)	k_{cat}/K_M ($M^{-1} s^{-1}$)
Aspartate–glyoxylate aminotransferase (BhcA)	Glyoxylate	58 ± 1	0.43 ± 0.02	1.34×10^5
	L-Aspartate	56 ± 1	2.51 ± 0.10	2.25×10^4
	Glycine	0.76 ± 0.01	9.52 ± 0.40	7.97×10^1
	Oxaloacetate	0.76 ± 0.02	2.90 ± 0.27	2.62×10^2
	L-Serine	8.8 ± 0.3	2.10 ± 0.24	4.20×10^3
	L-Glutamate	5.0 ± 0.3	20.62 ± 2.33	2.44×10^2
β-Hydroxyaspartate dehydratase (BhcB)	(2*R*, 3*S*)-β-Hydroxyaspartate	35 ± 1	0.20 ± 0.02	1.75×10^5
β-Hydroxyaspartate aldolase (BhcC)	Glyoxylate	86 ± 4	0.23 ± 0.03	3.72×10^5
	Glycine	91 ± 2	4.31 ± 0.34	2.11×10^4
	(2*R*, 3*S*)-β-Hydroxyaspartate	33 ± 1	0.28 ± 0.03	1.18×10^5
	D-Threonine	76 ± 2	9.24 ± 0.86	8.25×10^3
Iminosuccinate reductase (BhcD)	Iminosuccinate	201 ± 10	0.09 ± 0.01	2.29×10^6
	NADH	–	0.02 ± 0.003	–
	NADPH	–	0.33 ± 0.05	–

Data are mean ± s.d., as determined from nonlinear fits of 18 data points with GraphPad Prism 8. Michaelis–Menten fits of enzyme kinetics and an SDS–PAGE gel showing purified proteins are provided in Extended Data Fig. 4 and Supplementary Fig. 1, respectively. For BhcA, kinetics for glyoxylate and L-aspartate were measured with 20 mM L-aspartate and 5 mM glyoxylate, respectively, and kinetics for glycine and oxaloacetate were measured with 20 mM oxaloacetate and 30 mM glycine, respectively. Kinetics for L-serine and L-glutamate were measured with 5 mM glyoxylate. For BhcC, kinetics for glycine and glyoxylate were measured with 5 mM glyoxylate and 20 mM glycine, respectively.

The kinetic parameters of all enzymes of the BHAC are reported in Table 1. The complete reaction sequence of the pathway is shown in Fig. 1b. The cycle extends the originally proposed reaction sequence[5] by the IRED reaction. Overall, the BHAC converts two molecules of glyoxylate (C_2) into oxaloacetate (C_4) without the loss of carbon as CO_2, under consumption of just one reducing equivalent and regeneration of the catalytic amino donor, which makes it one of the most efficient glyoxylate assimilation pathways described to date (Supplementary Table 1). Oxaloacetate formed in the BHAC can directly enter the tricarboxylic acid cycle or serve as substrate for anabolic reactions. The pathway is essential for the growth of *P. denitrificans* in the presence of glycolate and glyoxylate, and its enzymes are highly expressed and active in cells grown in the presence of glycolate (Extended Data Fig. 5). Glyoxylate negatively affected the interaction of the transcriptional regulator BhcR with the promoter region of the bhc gene cluster (Extended Data Fig. 5).

We next studied the phylogenetic distribution of the BHAC. The bhc gene cluster is widespread among the Rhizobiales and Rhodobacterales orders of the Alphaproteobacteria, and is also found in several gammaproteobacterial orders (Extended Data Fig. 6). Most of these bacteria were isolated from marine habitats, and the *Roseobacter* group within the Rhodobacterales is one of the three major bacterial groups responding to phytoplankton blooms[27]; *Roseobacter*-group bacteria can constitute up to 15% of the bacterial community in these blooms[28]. Notably, 94% of the isolates with the bhc gene cluster also encode glycolate oxidase in their genomes, enabling them to oxidize glycolate to glyoxylate for subsequent assimilation by the BHAC (Supplementary Data 1). BhcC is also ubiquitously present in marine metagenomes collected on the *Tara* Oceans expedition (Extended Data Fig. 7 and Supplementary Data 2), suggesting that the BHAC functions in glycolate assimilation in marine environments worldwide. Notably, the BHAC (represented by BhcC) is on average 20-fold more abundant than the glycerate pathway (represented by Gcl) in these datasets (Extended Data Fig. 7d).

To investigate the ecological importance of the BHAC, we focused our analyses on Helgoland (Extended Data Fig. 8a, b), an island in the North Sea that has already been used extensively as a study site to investigate the succession of bacterial populations during algal blooms[28,29]. We analysed metagenomes from seawater samples collected between 2010 and 2012 at Helgoland and detected the bhc gene cluster in all years at intermediate abundances (up to 3 reads per kilobase per million reads (RPKM), corresponding to roughly 1.5% of all cells[28]) (Extended Data Fig. 8c–e and Supplementary Data 3). To further investigate the role of the BHAC in situ, we monitored the spring phytoplankton bloom at Helgoland from March to May 2018. We determined chlorophyll *a* (Chl *a*) fluorescence as proxy for phytoplankton biomass and total microbial cell counts for each working day. Glycolate concentrations in the seawater were determined weekly.

The 2018 spring bloom was dominated by pennate diatoms and consisted of two peaks in phytoplankton growth in late April and late May (Fig. 3a). We determined a background concentration of glycolate in the seawater of 300 nM before the bloom, which is in line with previous measurements[1,2,4,9–11,30–39] (600 ± 340 nM) (Extended Data Fig. 1). During the bloom, from early March to late May, glycolate concentrations increased by approximately 350 nM (Fig. 3b), indicating the accumulation of phytoplankton-derived glycolate. At three time points in April and May, before and during the algal bloom, we determined bulk uptake rates of glycolate in the sea water. Glycolate uptake rates were in line with previously reported values[12] and increased more than threefold from 1.46 nM h^{-1} to 4.68 nM h^{-1} between the first and the last measurement (Fig. 3b), indicating that the capacity for glycolate uptake had multiplied at the same factor as the total microbial cell counts. Notably, these rates are comparable to uptake and consumption rates for dimethylsulfoniopropionate in the open ocean[40] and would enable a turnover of the total glycolate pool at our sampling site every 5–10 days. The bhc gene cluster was prevalent during the progression of the phytoplankton bloom. *bhcC* genes were detected at all of the time points, with the highest abundance per cell (around 1.5%) during the peaks of the phytoplankton bloom in April and May (Fig. 3c, Extended Data Fig. 9 and Supplementary Data 4). Transcription of *bhcC* was confirmed before and during the spring bloom (Fig. 3d, e), indicating that the BHAC is an active route for glycolate assimilation in the ocean.

In summary, our study provides the full reaction sequence and genetic basis of the BHAC. We demonstrate the biochemistry of the pathway, which involves a previously unknown family of IREDs, and provide support for its ecological importance in the assimilation of phytoplankton-derived dissolved organic carbon. The discovery of the BHAC as a ubiquitous pathway in marine environments adds a new

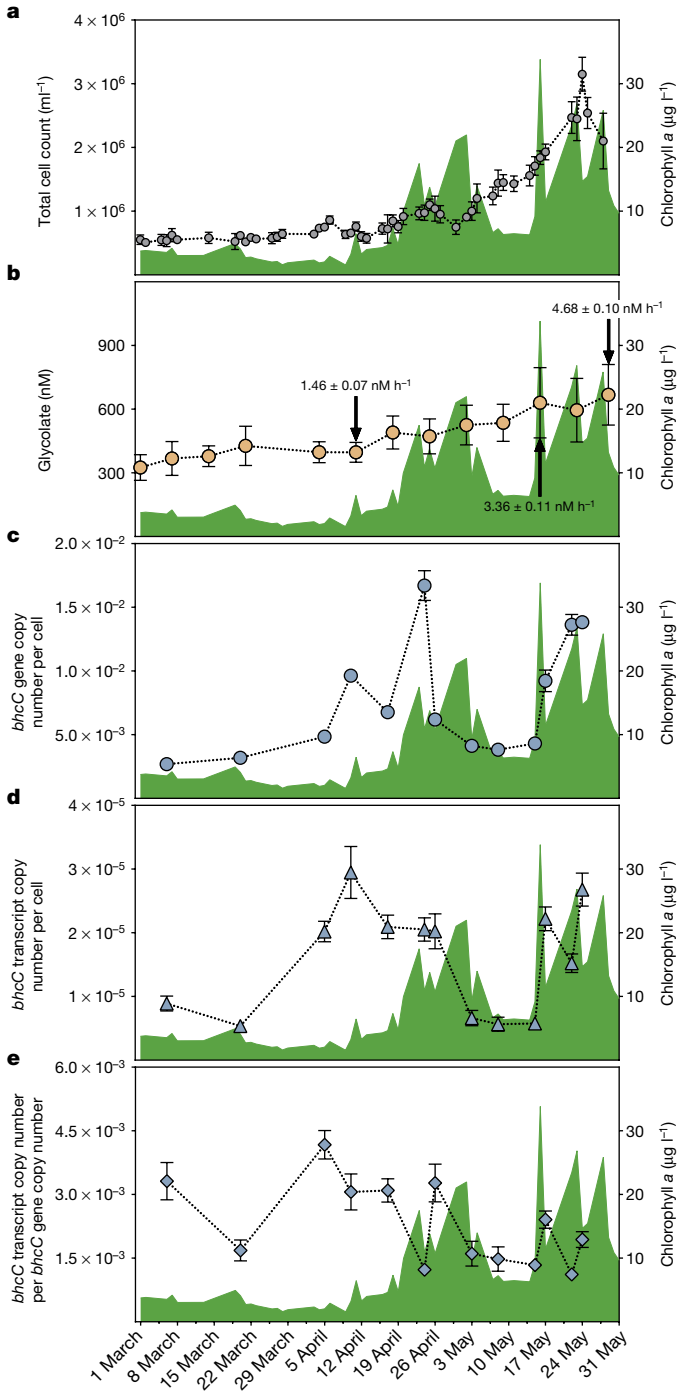


Fig. 3 | The BHAC during the spring phytoplankton bloom 2018 at Helgoland. a, From 1 March to 31 May, total microbial cell counts (grey) and Chl *a* concentrations (green) were determined each working day ($n = 1$). **b**, The concentration of glycolate (light brown) was determined once per week using liquid chromatography with mass spectrometry and increased from approximately 300 nM to around 650 nM. The uptake rate of bulk glycolate was determined at three time points through ^{14}C-glycolate incorporation and uptake rates are indicated. Data are the mean ± s.d. of $n = 5$ seawater samples for glycolate concentrations, and of $n = 4$ seawater samples for glycolate uptake rates. **c**, The *bhcC* gene copy number per cell (blue circles) was determined using qPCR. **d**, The *bhcC* transcript copy number per cell (blue triangles) was determined via cDNA synthesis followed by qPCR. **e**, *bhcC* transcript copy number divided by *bhcC* gene copy number (blue diamonds). **c**–**e**, Data are mean ± s.d.; $n = 3$ independent experiments.

dimension to the biochemical cycle of glycolate, an abundant organic acid in the global oceans. As the BHAC requires only one reducing equivalent and enables carbon-conserving glycolate assimilation, it may confer an advantage compared to the glycerate pathway, which releases CO_2. This may explain the high prevalence of the BHAC in marine Proteobacteria and could provide a starting point for future studies that investigate carbon fluxes from phytoplankton to heterotrophic bacterioplankton.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1748-4.

1. Wright, R. T. & Shah, N. M. Trophic role of glycolic acid in coastal seawater. II. Seasonal changes in concentration and heterotrophic use in Ipswich Bay, Massachusetts, USA. *Mar. Biol.* **43**, 257–263 (1977).
2. Wright, R. T. & Shah, N. M. Trophic role of glycolic acid in coastal seawater. I. Heterotrophic metabolism in seawater and bacterial cultures. *Mar. Biol.* **33**, 175–183 (1975).
3. Fogg, G. E. The ecological significance of extracellular products of phytoplankton photosynthesis. *Bot. Mar.* **26**, 3–14 (1983).
4. Lau, W. W., Keil, R. G. & Armbrust, E. V. Succession and diel transcriptional response of the glycolate-utilizing component of the bacterial community during a spring phytoplankton bloom. *Appl. Environ. Microbiol.* **73**, 2440–2450 (2007).
5. Kornberg, H. L. & Morris, J. G. β-Hydroxyaspartate pathway: a new route for biosyntheses from glyoxylate. *Nature* **197**, 456–457 (1963).
6. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998).
7. Duarte, C. M. & Cebrian, J. The fate of marine autotrophic production. *Limnol. Oceanogr.* **41**, 1758–1766 (1996).
8. Hellebust, J. A. Excretion of some organic compounds by marine phytoplankton. *Limnol. Oceanogr.* **10**, 192–206 (1965).
9. Tolbert, N. E. & Zill, L. P. Excretion of glycolic acid by algae during photosynthesis. *J. Biol. Chem.* **222**, 895–906 (1956).
10. Leboulanger, C., Descolasgros, C. & Jupin, H. HPLC determination of glycolic acid in seawater. An estimation of phytoplankton photorespiration in the Gulf of Lions, western Mediterranean Sea. *J. Plankton Res.* **16**, 897–903 (1994).
11. Leboulanger, C., Oriol, L., Jupin, H. & Descolas-Gros, C. Diel variability of glycolate in the eastern tropical Atlantic Ocean. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **44**, 2131–2139 (1997).
12. Casey, J. R., Ferrón, S. & Karl, D. M. Light-enhanced microbial organic carbon yield. *Front. Microbiol.* **8**, 2157 (2017).
13. Lau, W. W. & Armbrust, E. V. Detection of glycolate oxidase gene *glcD* diversity among cultured and environmental marine bacteria. *Environ. Microbiol.* **8**, 1688–1702 (2006).
14. Carini, P., Steindler, L., Beszteri, S. & Giovannoni, S. J. Nutrient requirements for growth of the extreme oligotroph 'Candidatus *Pelagibacter ubique*' HTCC1062 on a defined medium. *ISME J.* **7**, 592–602 (2013).
15. Eiler, A. et al. Tuning fresh: radiation through rewiring of central metabolism in streamlined bacteria. *ISME J.* **10**, 1902–1914 (2016).
16. Tripp, H. J. et al. Unique glycine-activated riboswitch linked to glycine–serine auxotrophy in SAR11. *Environ. Microbiol.* **11**, 230–238 (2009).
17. Krakow, G. & Barkulis, S. S. Conversion of glyoxylate to hydroxypyruvate by extracts of *Escherichia coli*. *Biochim. Biophys. Acta* **21**, 593–594 (1956).
18. Hansen, R. W. & Hayashi, J. A. Glycolate metabolism in *Escherichia coli*. *J. Bacteriol.* **83**, 679–687 (1962).
19. Kornberg, H. L. & Morris, J. G. The utilization of glycollate by *Micrococcus denitrificans*: the β-hydroxyaspartate pathway. *Biochem. J.* **95**, 577–586 (1965).
20. Gibbs, R. G. & Morris, J. G. Assay and properties of β-hydroxyaspartate aldolase from *Micrococcus denitrificans*. *Biochim. Biophys. Acta* **85**, 501–503 (1964).
21. Gibbs, R. G. & Morris, J. G. Purification and properties of *erythro*-β-hydroxyaspartate dehydratase from *Micrococcus denitrificans*. *Biochem. J.* **97**, 547–554 (1965).
22. Liu, J. Q., Dairi, T., Itoh, N., Kataoka, M. & Shimizu, S. A novel enzyme, d-3-hydroxyaspartate aldolase from *Paracoccus denitrificans* IFO 13301: purification, characterization, and gene cloning. *Appl. Microbiol. Biotechnol.* **62**, 53–60 (2003).
23. Mortarino, M. et al. l-Aspartate oxidase from *Escherichia coli*. I. Characterization of coenzyme binding and product inhibition. *Eur. J. Biochem.* **239**, 418–426 (1996).
24. Lenz, M., Borlinghaus, N., Weinmann, L. & Nestl, B. M. Recent advances in imine reductase-catalyzed reactions. *World J. Microbiol. Biotechnol.* **33**, 199 (2017).
25. Hochreiter, M. C. & Schellenberg, K. A. α-Iminoglutarate formation by beef liver l-glutamate dehydrogenase. Detection by borohydride or dithionite reduction to glutamate. *J. Am. Chem. Soc.* **91**, 6530–6531 (1969).
26. Hallen, A., Cooper, A. J., Smith, J. R., Jamie, J. F. & Karuso, P. Ketimine reductase/CRYM catalyzes reductive alkylamination of α-keto acids, confirming its function as an imine reductase. *Amino Acids* **47**, 2457–2461 (2015).

27. Buchan, A., LeCleir, G. R., Gulvik, C. A. & González, J. M. Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nat. Rev. Microbiol*. **12**, 686–698 (2014).
28. Teeling, H. et al. Recurring patterns in bacterioplankton dynamics during coastal spring algae blooms. *eLife* **5**, e11888 (2016).
29. Teeling, H. et al. Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. *Science* **336**, 608–611 (2012).
30. Shah, N. M. & Wright, R. T. Occurrence of glycolic acid in coastal sea-water. *Mar. Biol*. **24**, 121–124 (1974).
31. Fogg, G. E., Burton, N. F. & Coughlan, S. J. The occurrence of glycollic acid in Antarctic waters. *Br. Antarct. Surv. Bull*. **41 & 42**, 193–195 (1975).
32. Hasan-Al, R. H., Coughlan, S. J., Pant, A. & Fogg, G. E. Seasonal variations in phytoplankton and glycollate concentrations in the Menai Straits, Anglesey. *J. Mar. Biol. Assoc. U.K*. **55**, 557–565 (1975).
33. Edenborn, H. M. & Litchfield, C. D. Glycolate turnover in the water column of the New York Bight apex. *Mar. Biol*. **95**, 459–467 (1987).
34. Leboulanger, C., Serve, L., Comellas, L. & Jupin, H. Determination of glycolic acid released from marine phytoplankton by post-derivatization gas chromatography mass spectrometry. *Phytochem. Anal*. **9**, 5–9 (1998).
35. Lord, J. M., Codd, G. A. & Merrett, M. J. The effect of light quality on glycolate formation and excretion in algae. *Plant Physiol*. **46**, 855–856 (1970).
36. Smith, W. O. Extracellular release of glycolic acid by a marine diatom. *J. Phycol*. **10**, 30–33 (1974).
37. Leboulanger, C., Martin-Jezequel, V., Descolas-Gros, C., Sciandra, A. & Jupin, H. J. Photorespiration in continuous culture of *Dunaliella tertiolecta* (Chlorophyta): relationships between serine, glycine, and extracellular glycolate. *J. Phycol*. **34**, 651–654 (1998).
38. Schnitzler Parker, M., Armbrust, E. V., Piovia-Scott, J. & Keil, R. G. Induction of photorespiration by light in the centric diatom *Thalassiosira weissflogii* (*Bacillariophyceae*): molecular characterization and physiological consequences. *J. Phycol*. **40**, 557–567 (2004).
39. Bertilsson, S., Berglund, O., Pullin, M. J. & Chisholm, S. W. Release of dissolved organic matter by *Prochlorococcus*. *Vie Milieu* **55**, 225–231 (2005).
40. Simó, R. & Pedrós-Alió, C. Short-term variability in the open ocean cycle of dimethylsulfide. *Glob. Biogeochem. Cycles* **13**, 1173–1181 (1999).

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Chemicals and reagents

Unless otherwise stated, all chemicals and reagents were acquired from Sigma-Aldrich and were of the highest purity available.

Strains, medium and cultivation conditions

All strains used in this study are listed in Supplementary Table 2. *Escherichia coli* TOP10 (for genetic work), ST18 (for plasmid conjugation) and BL21 AI (for protein expression) were grown at 37 °C in lysogeny broth (LB)[41].

P. denitrificans DSM 413[42] and its derivatives were grown at 30 °C in LB or in mineral salt medium with TE3-Zn trace elements[43] supplemented with various carbon sources. To monitor growth, the optical density at 600 nm (OD_{600}) of culture samples was determined on a photospectrometer (Merck Chemicals).

Vector construction

The genes encoding the four enzymes of the BHAC (*bhcABCD*) as well as the *bhcR* gene encoding the transcriptional regulator were cloned into the standard expression vector pET16b (Merck Chemicals). To this end, the respective genes were amplified from genomic DNA of *P. denitrificans* DSM 413 with the primers provided in Supplementary Table 3. The resulting PCR products were digested with suitable restriction endonucleases (Thermo Fisher Scientific) as given in Supplementary Table 3 and ligated into the expression vector pET16b that had been digested with the same enzymes to create a vector for heterologous expression of the respective protein. Successful cloning of the desired open reading frames was verified by DNA sequencing (Eurofins Genomics). All plasmids used in this study are listed in Supplementary Table 2.

Expression and purification of recombinant proteins

For heterologous overexpression of the BhcA, BhcB, BhcC and BhcD enzymes, the corresponding plasmid encoding the respective enzyme was first transformed into chemically competent *E. coli* BL21 AI cells. The cells were then grown on LB agar plates containing 100 μg ml^{-1} ampicillin at 37 °C overnight. A starter culture in selective LB medium was inoculated from a single colony on the next day and left to grow overnight at 37 °C in a shaking incubator. The starter culture was used on the next day to inoculate an expression culture in selective terrific broth (TB) medium at a 1:100 dilution. The expression culture was grown at 37 °C in a shaking incubator to an OD_{600} of 0.5–0.7, induced with 0.5 mM isopropyl-β-D-thiogalactoside (IPTG) and 0.2% L-arabinose and subsequently grown overnight at 18 °C in a shaking incubator. Cells were collected at 6,000*g* for 15 min at 4 °C and cell pellets were stored at −20 °C until purification of enzymes. Cell pellets were resuspended in twice their volume in buffer A (300 mM NaCl, 25 mM Tris-HCl pH 8.0, 15 mM imidazole, 1 mM β-mercaptoethanol, 0.1 mM $MgCl_2$, 0.01 mM PLP and one tablet of SIGMAFAST protease inhibitor cocktail, EDTA-free per litre). The cell suspension was treated with a Sonopuls GM200 sonicator (BANDELIN Electronic) at an amplitude of 50% to lyse the cells and subsequently centrifuged at 50,000*g* and 4 °C for 1 h. The filtered supernatant (0.45-μm filter; Sarstedt) was loaded onto Protino Ni-NTA Agarose (Macherey-Nagel) in a gravity column, which had previously been equilibrated with 5 column volumes of buffer A. The column was washed with 20 column volumes of buffer A and 5 column volumes of 85% buffer A and 15% buffer B and the His-tagged protein was eluted with buffer B (buffer A with 500 mM imidazole). The eluate was desalted using PD-10 desalting columns (GE Healthcare) and buffer C (100 mM NaCl, 25 mM Tris-HCl pH 8.0, 1 mM $MgCl_2$, 0.01 mM PLP, 0.1 mM dithiothreitol (DTT)). This was followed by purification on a size-exclusion column (Superdex 200 pg, HiLoad 16/600; GE Healthcare) connected to an ÄKTA Pure system (GE Healthcare) using buffer C. The concentrated protein solution (2 ml) was injected, and the flow was kept constant at 1 ml min^{-1}. Elution fractions containing pure protein were determined via SDS–PAGE analysis[44] on 12.5% gels. Purified enzymes in buffer C were used for crystallization or stored at −20 °C in buffer C containing 50% glycerol for later use in enzymatic assays.

BhcR was expressed and purified in the same way, except that buffer A contained 100 mM KCl, 20 mM HEPES-KOH pH 7.5, 10 mM $MgCl_2$, 4 mM β-mercaptoethanol, 5% glycerol and one tablet of SIGMAFAST protease inhibitor cocktail, EDTA-free per litre. Buffer C contained 100 mM KCl, 20 mM HEPES-KOH pH 7.5, 10 mM $MgCl_2$, 5% glycerol and 1 mM DTT.

NADH-dependent malate dehydrogenase (Mdh) and NADPH-dependent glyoxylate reductase (GhrA) from *E. coli* were overexpressed using the respective strains from the ASKA collection[45]. A starter culture in selective LB medium (34 μg ml^{-1} chloramphenicol) was inoculated from a single colony and left to grow overnight at 37 °C in a shaking incubator. The starter culture was used on the next day to inoculate an expression culture in selective TB medium at a 1:100 dilution. The expression culture was grown at 37 °C in a shaking incubator to an OD_{600} of 0.6, induced with 0.5 mM IPTG and grown another 4 h at 37 °C in a shaking incubator. The enzymes were affinity-purified in the same way as described above, except that buffer A contained 200 mM NaCl, 50 mM potassium phosphate pH 7.0, 15 mM imidazole, 1 mM β-mercaptoethanol and one tablet of SIGMAFAST protease inhibitor cocktail, EDTA-free per litre. Buffer C contained 100 mM NaCl, 50 mM potassium phosphate pH 7.0 and 0.1 mM DTT. The purified enzyme was stored at −20 °C in buffer C containing 50% glycerol.

Enzyme activity assays

For all enzyme assays, the oxidation of NADH or NADPH was followed at 340 nm or 360 nm on a Cary 60 UV-Vis photospectrometer (Agilent) in quartz cuvettes with a path length of 1 mm or 10 mm (Hellma Optik).

The enzyme assay to determine the kinetic parameters of BhcA with glyoxylate and L-aspartate as substrates was performed at 30 °C in a total volume of 300 μl. The reaction mixture contained 100 mM potassium phosphate buffer pH 7.5, 0.1 mM PLP, 0.2 mM NADH, different amounts of the respective substrates and 32 nM BhcA. Five hundred nanomolar Mdh was added as a coupling enzyme to convert oxaloacetate into malate. Kinetics for glyoxylate were measured with 20 mM L-aspartate; kinetics for L-aspartate were measured with 5 mM glyoxylate. To determine the kinetic parameters with oxaloacetate and glycine as substrates, the same assay mixture was used and 3 μM GhrA was added as a coupling enzyme to convert glyoxylate into glycolate. Kinetics for glycine were measured with 20 mM oxaloacetate; kinetics for oxaloacetate were measured with 30 mM glycine. To determine the kinetic parameters with L-serine or L-glutamate and glyoxylate as substrates, the same assay mixture was used and BhcB (3 μM), BhcC (1 μM) and Mdh (500 nM) were added as coupling enzymes. Kinetics for L-serine and L-glutamate were measured with 5 mM glyoxylate.

The enzyme assay to determine the kinetic parameters of BhcB was performed at 30 °C in a total volume of 300 μl. The reaction mixture contained 100 mM potassium phosphate buffer pH 7.5, 0.1 mM PLP, 0.2 mM NADH, different amounts of the substrate (2*R*, 3*S*)-β-hydroxyaspartate and 29 nM BhcB. Five hundred and eighty nanomolar BhcD was added as a coupling enzyme to convert iminosuccinate into L-aspartate. (2*R*, 3*S*)-β-Hydroxyaspartate was custom-synthesized by NewChem and was determined to be >95% pure by NMR analysis.

The enzyme assay to determine the kinetic parameters of BhcC with glyoxylate and glycine as substrates was performed at 30 °C in a total volume of 300 μl. The reaction mixture contained 100 mM potassium phosphate buffer pH 7.5, 0.1 mM PLP, 0.2 mM NADH, 0.5 mM $MgCl_2$, different amounts of the respective substrates and 4 nM BhcC. BhcB (200 nM) and BhcD (2 μM) were added as coupling enzymes. Kinetics

for glycine were measured with 5 mM glyoxylate; kinetics for glyoxylate were measured with 20 mM glycine. To determine the kinetic parameters with (2*R*, 3*S*)-β-hydroxyaspartate as substrate, the same assay mixture was used and 3 μM GhrA was added as a coupling enzyme to convert glyoxylate into glycolate. To determine the kinetic parameters with D-threonine as substrate, the same assay mixture was used and 3 μM alcohol dehydrogenase from *Saccharomyces cerevisiae* (Sigma-Aldrich) was added as coupling enzyme to convert acetaldehyde into ethanol.

The enzyme assay to determine the apparent kinetic parameters of BhcD was performed at 30 °C in a total volume of 250 μl. The reaction mixture contained 100 mM potassium phosphate buffer pH 7.5, 0.2 mM NADH, 0.1 mM PLP, different amounts of (2*R*, 3*S*)-β-hydroxyaspartate, and appropriate amounts of the enzymes BhcB and BhcD. Kinetics for iminosuccinate were measured with 15 nM BhcD, different amounts of (2*R*, 3*S*)-β-hydroxyaspartate and BhcB. To a given amount of (2*R*, 3*S*)-β-hydroxyaspartate, a tenfold molar excess of BhcB was added to start the reaction and completely and almost instantly convert the substrate pool into iminosuccinate. The initial reaction velocity of BhcD was determined after a mixing period of 3 s and the apparent concentration of iminosuccinate at this point in time was calculated on the basis of previously published values[23]. Kinetics for NADH and NADPH were measured with 2 mM (2*R*, 3*S*)-β-hydroxyaspartate, 214 nM BhcB, 28 nM BhcD and different amounts of the respective cofactor. No activity was measurable in a reaction mixture containing 100 mM potassium phosphate buffer pH 7.5, 0.2 mM NADH, 0.1 mM PLP and 3 mM oxaloacetate as well as 9 mM ammonium as putative substrates for BhcD.

The enzyme assay to generate iminosuccinate from (2*R*, 3*S*)-β-hydroxyaspartate (catalysed by BhcB) and further chemical reduction of iminosuccinate to L-aspartate with the reducing agent $NaBH_3CN$[46] was performed at 30 °C in a total volume of 1 ml. The reaction mixture contained 50 mM Tris pH 7.5, 1 mM (2*R*, 3*S*)-β-hydroxyaspartate, 0.1 mM PLP, 1 mM $MgCl_2$, 214 nM BhcB and 1 mM $NaBH_3CN$. The reaction was carried out in D_2O. Aliquots of 180 μl were taken after 0, 0.5, 1, 2 and 3 min and the reaction was immediately stopped by quenching with formic acid (4% final concentration). The samples were centrifuged at 17,000*g* and 4 °C for 15 min and the supernatant diluted 1:4 in double-distilled water for liquid chromatography–mass spectrometry (LC–MS) analysis. In negative control experiments, $NaBH_3CN$ was omitted from the reaction mixture. The same experiment was performed with added BhcD instead of $NaBH_3CN$ to enzymatically reduce iminosuccinate to L-aspartate. The reaction mixture contained 50 mM Tris pH 7.5, 1 mM (2*R*, 3*S*)-β-hydroxyaspartate, 2 mM NADH, 0.1 mM PLP, 1 mM $MgCl_2$, 214 nM BhcB and 28 nM BhcD.

LC–MS measurements were performed using an Agilent 6550 iFunnel Q-TOF LC–MS system equipped with an electrospray ionization (ESI) source set to negative ionization mode. LC was carried out as follows. The analytes were separated on an aminopropyl column (30 mm × 2 mm, particle size 3 μm, 100 Å; Luna NH2, Phenomenex) using a mobile phase system consisting of 95:5 20 mM ammonium acetate pH 9.3 (adjusted with ammonium hydroxide to a final concentration of approximately 10 mM): acetonitrile (A) and acetonitrile (B). Chromatographic separation was carried out using the following gradient condition at a flow rate of 250 μl min^{-1}: 0 min, 85% B; 3.5 min, 0% B, 7 min, 0% B; 7.5 min, 85% B; 8 min, 85% B. Column oven and autosampler temperature were maintained at 15 °C. The ESI source was set to the following parameters: capillary voltage was set at 3.5 kV and nitrogen gas was used as nebulizing (20 psig), drying (13 l min^{-1}, 225 °C) and sheath gas (12 l min^{-1}, 400 °C). The Q-TOF mass detector was calibrated before measurement using an ESI-L Low Concentration Tuning Mix (Agilent) with residuals and corrected residuals less than 2 ppm and 1 ppm, respectively. MS data were acquired with a scan range of 50–600 *m*/*z*. Autorecalibration was carried out using 113 *m*/*z* as reference mass. Subsequent peak integration of all analytes was performed using eMZed 2.29.4.0[47].

Enzyme activity assays in *P. denitrificans* cell extracts

P. denitrificans cultures were collected during mid-exponential phase (OD_{600} of 0.5–0.7), resuspended in ice-cold 100 mM potassium phosphate buffer (pH 7.2) and lysed by sonication. Cell debris was separated by centrifugation at 35,000*g* and 4 °C for 1 h. Total protein concentrations of the resulting cell-free extracts were determined by Bradford assay[48] using bovine serum albumin as standard. The assays for activity of BhcABCD were performed as described above, except that 100 mM potassium phosphate buffer pH 7.5 was replaced with 100 mM Tris pH 7.5. During BhcD assays, 90 μl samples were taken after 0.5, 1 and 2 min, and the reaction was immediately stopped by quenching with formic acid (4% final concentration). The samples were centrifuged at 17,000*g* and 4 °C for 15 min and the supernatant diluted 1:10 in double-distilled water for LC–MS analysis. L-Malate and L-aspartate in the samples were quantified using a standard curve of each compound ranging from 10 μM to 1,000 μM.

Genetic modification of *P. denitrificans*

The upstream and downstream flanking regions of the *bhcABCD* genes from *P. denitrificans* DSM 413 were cloned into the gene deletion vector pREDSIX[49]. To this end, the flanking regions were amplified from genomic DNA of *P. denitrificans* DSM 413 with the primers given in Supplementary Table 3. The resulting PCR products were used to perform Gibson assembly with the vector pREDSIX, which had been digested with MfeI. Subsequently, the resulting vector was digested with NdeI, and a kanamycin-resistance cassette, which had been cut out of the vector pRGD-Kan with NdeI, was ligated into the cut site to generate the final vectors for gene deletion. For gene deletion of each of the genes *bhcABCD* separately and of the complete bhc gene cluster, the corresponding plasmid was first transformed into chemically competent *E. coli* ST18[50] cells, which were then grown on LB agar plates containing 100 μg ml^{-1} ampicillin, 50 μg ml^{-1} kanamycin and 50 μg ml^{-1} aminolevulinic acid at 37 °C overnight. A culture in selective LB medium was inoculated the next day and left to grow overnight at 37 °C. The cultures were diluted the next morning to an OD_{600} of 0.1. A culture of wild-type *P. denitrificans* DSM 413 in LB medium was inoculated from a glycerol stock and grown at 30 °C. ST18 cultures were collected at an OD_{600} of around 0.7, and the *P. denitrificans* culture was collected at an OD_{600} of about 1.3. All cell pellets were washed once with sterile 10 mM $MgSO_4$ and resuspended to an OD_{600} of approximately 10 in sterile 10 mM $MgSO_4$. Suspensions of ST18 cells and *P. denitrificans* cells were mixed in a 2:1 ratio and spotted on minimal medium agar plates without any carbon source. Plates were incubated at 30 °C overnight. The next morning, spots were removed from the plates, resuspended in LB medium and plated on LB agar plates containing 25 μg ml^{-1} kanamycin. Plates were incubated at 30 °C for 3 days. The respective gene deletion was verified by colony PCR and DNA sequencing (Eurofins Genomics) and the deletion strain was propagated in selective LB medium.

High-throughput growth assays with *P. denitrificans* strains

Cultures of wild-type *P. denitrificans* DSM 413 and gene deletion strains were pre-grown at 30 °C in LB medium containing 25 μg ml^{-1} kanamycin, when necessary. Cells were collected, washed once with minimal medium containing no carbon source and used to inoculate growth cultures of 180 μl minimal medium containing an appropriate carbon source as well as 25 μg ml^{-1} kanamycin for gene deletion strains. Growth in 96-well plates (Thermo Fisher Scientific) was monitored at 30 °C at 600 nm in a Tecan Infinite M200Pro reader (Tecan). The resulting data were evaluated using GraphPad Prism 8.0.0.

Whole-cell shotgun proteomics

To acquire the proteome of *P. denitrificans* growing on different carbon sources, 30 ml cultures were grown to mid-exponential phase (OD_{600} of around 0.4) in minimal medium supplemented with 30 mM succinate

or 60 mM glycolate. Four replicate cultures were grown for each carbon source. Main cultures were inoculated from precultures grown in the same medium at a dilution of 1:1,000. Cultures were collected by centrifugation at 4,000*g* and 4 °C for 15 min. The supernatant was discarded and pellets were washed in 40 ml phosphate-buffered saline (PBS; 137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4, pH 7.4). After washing, cell pellets were resuspended in 1 ml PBS, transferred into Eppendorf tubes and centrifuged as described above. Cell pellets in Eppendorf tubes were snap-frozen in liquid nitrogen and stored at −80 °C until they were used for the preparation of samples for LC–MS analysis and label-free quantification.

For protein extraction, bacterial cell pellets were resuspended in 4% SDS and lysed by heating (95 °C, 15 min) and sonication (Hielscher Ultrasonics). Reduction was performed for 15 min at 90 °C in the presence of 5 mM tris(2-carboxyethyl)phosphine followed by alkylation using 10 mM iodoacetamide at 25 °C for 30 min. The protein concentration in each sample was determined using the BCA protein assay kit (Thermo Fisher Scientific) following the manufacturer's instructions. Protein clean-up and tryptic digestion were performed using the SP3 protocol as previously described[51] with minor modifications regarding protein digestion temperature and solid-phase extraction of peptides. SP3 beads were obtained from GE Healthcare. Trypsin (1 µg, Promega) was used to digest 50 µg of total solubilized protein from each sample. Tryptic digestion was performed overnight at 30 °C. Subsequently, all protein digestions were desalted using C18 microspin columns (Harvard Apparatus) according to the manufacturer's instructions.

LC–MS/MS analysis of protein digestions was performed on a Q-Exactive Plus mass spectrometer connected to an electrospray ion source (Thermo Fisher Scientific). Peptide separation was carried out using an Ultimate 3000 nanoLC-system (Thermo Fisher Scientific), equipped with an in-house-packed C18 resin column (Magic C18 AQ 2.4 µm; Dr. Maisch). The peptides were first loaded onto a C18 precolumn (preconcentration set-up) and then eluted in backflush mode with a gradient from 98% solvent A (0.15% formic acid) and 2% solvent B (99.85% acetonitrile and 0.15% formic acid) to 25% solvent B over 105 min, continued from 25% to 35% of solvent B up to 135 min. The flow rate was set to 300 nl min^{-1}. The data acquisition mode for the initial label-free quantification study was set to obtain one high-resolution MS scan at a resolution of 60,000 (*m/z* 200) with a scanning range from 375 to 1,500 *m/z* followed by MS/MS scans of the 10 most intense ions. To increase the efficiency of MS/MS shots, the charged-state screening modus was adjusted to exclude unassigned and singly charged ions. The dynamic exclusion duration was set to 30 s. The ion accumulation time was set to 50 ms (both MS and MS/MS). The automatic gain control was set to 3×10^6 for MS survey scans and 1×10^5 for MS/MS scans. Label-free quantification was performed using Progenesis QI (v.2.0). MS raw files were imported into Progenesis and the output data (MS/MS spectra) were exported in MGF format. MS/MS spectra were then searched using MASCOT (v.2.5) against a database of the predicted proteome from *P. denitrificans* downloaded from the UniProt database (https://www.uniprot.org/; download date 26 January 2017), containing 386 common contaminant and background proteins that were manually added. The following search parameters were used: full tryptic specificity required (cleavage after lysine or arginine residues); two missed cleavages allowed; carbamidomethylation (C) set as a fixed modification; and oxidation (M) set as a variable modification. The mass tolerance was set to 10 ppm for precursor ions and 0.02 Da for fragment ions for high-energy collision dissociation. Results from the database search were imported back into Progenesis, mapping peptide identifications to MS1 features. The peak heights of all MS1 features annotated with the same peptide sequence were summed, and protein abundance was calculated per LC–MS run. Next, the data obtained from Progenesis were evaluated using the SafeQuant R package v.2.2.2[52]. Then, the 1% false-discovery rate of identification and quantification as well as the intensity-based absolute quantification values were calculated.

Electrophoretic mobility shift assays

Fluorescently labelled DNA fragments for electrophoretic mobility shift assays were generated by PCR from genomic DNA of *P. denitrificans* DSM 413. For the *Pbhc* regulatory region, primers Pbhc_fw and Pbhc_rev-dye were used to generate a 238-bp fragment containing the putative *Pbhc* promoter. The primers bhcA_fw and bhcA_rev-dye were used to generate a 255-bp fragment containing a fragment of the *bhcA* gene as negative control. The primers *Pbhc*_rev-dye and *bhcA*_rev-dye were 5′-labelled with the Dyomics 781 fluorescent dye (Microsynth). Binding reactions between the DNA fragments (0.025 pmol), various amounts of the purified protein BhcR (400×, 2,000×, 4,000×, 10,000×, 20,000× and 40,000× molar excess), and various concentrations of glyoxylate (0.01, 0.05, 0.1, 0.2, 0.5 and 1 mM final concentration) were performed in buffer A (20 mM potassium phosphate pH 7.0, 1 mM DTT, 5 mM $MgCl_2$, 50 mM KCl, 15 µg ml^{-1} bovine serum albumin, 50 µg ml^{-1} herring sperm DNA, 5% v/v glycerol, 0.1% Tween-20) in a total volume of 20 µl. After the reaction mixtures were incubated at 37 °C for 20 min, the samples were loaded onto a native 5% polyacrylamide gel and electrophoretically separated at 110 V for 60 min. BhcR–DNA interactions were detected using an Odyssey FC Imaging System (LI-COR Biosciences).

Crystallization and structure determination of BhcC and BhcD

The sitting-drop vapour-diffusion method was used for crystallization at 16 °C. Purified BhcC (10 mg ml^{-1}) was mixed in a 1:1 ratio with solution A containing 20% PEG 3350, 0.2 M ammonium chloride, pH 6.3 (final drop volume 1.4 µl). Reservoirs were filled with 40 µl solution A. Crystals appeared within 14 days. Crystals were briefly soaked in mother liquor supplemented with 40% glycerol for cryoprotection before freezing in liquid nitrogen.

Purified BhcD (10 mg ml^{-1}) was mixed in a 1:1 ratio with solution B containing 20% PEG 3350, 0.2 M $Mg(NO_3)_2$, 5 mM NAD^+ and 5 mM Tb-Xo4, pH 6.4 (final drop volume 4 µl). Various additives were tested to improve crystal quality and size. The best results were achieved with the recently described nucleating and phasing agent Tb-Xo4[53]. Reservoirs were filled with 114 µl of solution B. Crystals appeared within a week. Crystals were briefly soaked in mother liquor supplemented with 40% ethylene glycol for cryoprotection before freezing in liquid nitrogen.

X-ray diffraction data were collected at the beamlines ID29 and ID30B of the ESRF (Grenoble, France) and at beamline P13 of DESY (Hamburg, Germany). The data were processed with the XDS[54] (build 20180126) and CCP4 v.7.0 software packages[55]. The structures were solved by molecular replacement. For BhcC, the structure of a D-threonine aldolase (PDB 4V15)[56] served as search model. For BhcD, a homology model was made based on the structure of L-alanine dehydrogenase (PDB 1OMO)[57] using Swiss-Model[58]. This homology model was then used as search model for the molecular replacement. The molecular replacement was carried out using Phaser of the Phenix software package[59] (v.1.14), built with Phenix.Autobuild and refined with Phenix.Refine. Additional modelling, manual refining and ligand fitting was done in Coot[60] (v.0.8.9). Final positional and B-factor refinements, as well as water picking, were performed using Phenix.Refine. The structure models for BhcC and BhcD were deposited at the Protein Data Bank in Europe (PDBe) under PDB accession numbers 6QKB and 6RQA, respectively. Figures were made using Pymol 1.8.

Analysis of North Sea metagenome data

Searches for the bhc gene cluster in 38 assembled surface seawater metagenomes sampled at the island of Helgoland between 2010 and 2012 were performed using the *Ruegeria pomeroyi* DSS-3 bhc gene cluster proteins as reference (NCBI protein IDs WP_011241924.1 (BhcR), WP_011241925.1 (BhcA), WP_011241926.1 (BhcB), WP_011241927.1 (BhcC), WP_044029519.1 (BhcD)). All identified proteins of the 38 metagenomes were searched against these proteins using DIAMOND[61] BLASTp and post-filtered to those hits for which the entire gene cluster

could be detected on a metagenome contig. These contigs were, if possible, linked to metagenome-assembled genomes (MAGs) binned from the same 38 metagenomes. MAGs were binned as previously described[62] and both the metagenome assemblies and MAGs are accessible under accession PRJEB28156 at the European Nucleotide Archive (ENA). MAG quality was assessed using CheckM v.1.0.7[63]. Abundance estimates of MAGs and the single unbinned contig were calculated based on read mapping as reads per kilobase per million reads (RPKM; 2 RPKM ≈ 1% relative abundance detected by fluorescence in situ hybridization[28]). Read mapping of all 38 metagenomes to MAGs and the single unbinned contig was performed as previously described[62] using BBMap v.35.14 (http://bbtools.jgi.doe.gov).

Phylogenetic analyses

A genome tree of bacterial strains and five MAGs with the bhc gene cluster was calculated using GTDBtk v.0.1.3 with GTDB v.86[64]. GTDBtk uses an alignment of 120 bacterial marker genes to infer taxonomic relationships. The GTDBtk calculated tree was subsampled to the 264 bhc gene cluster containing bacterial strains and the MAGs and visualized using iTOL[65].

Sequences of BhcABCD from 264 bacterial isolates and 6 metagenome contigs (five of which were linked to MAGs) were aligned using MUSCLE[66], manually curated to remove gaps and concatenated. A phylogenetic tree of concatenated sequences of BhcABCD was calculated using raxmlGUI[67] 1.5b2 using the PROTGAMMA model with Le-Gascuel substitution matrix[68], 100 bootstraps and 100 maximum-likelihood resamplings. The resulting tree was visualized using iTOL.

Sequences from the ornithine cyclodeaminase/μ-crystalline superfamily (Conserved Domain accession cl27428) and the type III PLP-dependent enzymes superfamily (Conserved Domain accession cl00261) were downloaded from the NCBI protein database and aligned using MUSCLE. Phylogenetic trees of the aligned sequences were calculated with raxmlGUI 1.5b2 using the PROTGAMMA model with Le-Gascuel substitution matrix, 100 bootstraps and 100 maximum-likelihood resamplings. The resulting trees were visualized using iTOL.

In total, 1,614 protein sequences from the ornithine cyclodeaminase/μ-crystalline superfamily were used for generation of a sequence similarity network (SSN) using the EFI-EST web tool[69] with a cut-off value of 1×10^{-50}. In this SSN, all connected sequences that shared 80% or more identity were grouped into a single node, resulting in 619 meta nodes. The SSN was visualized with Cytoscape 3.7.1 (https://cytoscape.org) and edges between nodes with less than 50% identity were removed.

Analysis of *Tara* Oceans metagenomes

BhcC from *P. denitrificans* DSM 413 (Uniprot A1B8Z1) and Gcl from *Starkeya novella* DSM 506 (Uniprot D7A6R1) were used as queries to search the OM-RGC_v1 database using the Ocean Gene Atlas[70] web tool (http://tara-oceans.mio.osupytheas.fr/ocean-gene-atlas) with a cut-off value of 1×10^{-100}. The resulting hits were inspected and sequences that were deemed to not belong to BhcC or Gcl were removed. The following criteria were used: at least 50% of the query sequence covered; at least two of the three residues A160, A195, S313 present for BhcC sequences; residues V25, V51, L421, L476, L478, I479[71] present for Gcl sequences. The coordinates of sampling sites with positive hits for BhcC in samples from surface water (0.22–3-μm size fraction) were downloaded and visualized using Ocean Data View 5.1.5 (Schlitzer, R., Ocean Data View, odv.awi.de, 2018). Taxonomic assignments of BhcC and Gcl sequences were downloaded and manually converted to GTDB taxonomy. Sequence IDs are listed in Supplementary Data 2.

Environmental sample collection and processing

Sampling was carried out on each working day (Monday–Friday) with the RV *Aade* (https://www.awi.de/en/expedition/ships/more-ships.html) at the research site 'Kabeltonne' (54° 11.3′ N, 7° 54.0′ E) from approximately 1 m water depth in 20 l carboys. The water samples for microbial biomass were subjected to fractionating filtration directly upon arrival in the Biologische Anstalt Helgoland laboratory (typically less than one hour after sampling). Three membrane 142-mm diameter filtration units were operated in parallel to keep filtration times to a minimum. First, samples were pre-filtered through 142-mm diameter 10-μm-pore-size polycarbonate filters (Merck Chemicals) by means of an air-pressure pump to remove large particles and eukaryotic plankton. Then, the water samples were filtered with air-pressure pumps onto 142-mm diameter 3-μm-pore-size polycarbonate filters (Merck Chemicals) to collect predominantly bacteria associated with smaller particles and algae. Afterwards, dedicated aliquots were filtered onto 142-mm diameter 0.2-μm pore-size-polyethersulfone filters (Merck Chemicals) for DNA and RNA extraction. Bacterioplankton dominated this 0.2-μm fraction. The entire filtration process for all fractions was usually finished within 3 h, that is, latest 4 h after the sampling. All filters were stored at −80 °C until further analyses.

Total cell counts

Samples were fixed with 1% formaldehyde and filtered onto polycarbonate membrane filters as described above. Total cell counts were determined from 10 ml fixed seawater samples. One filter section was cut and stained with 4′,6-diamidino-2-phenylindole (DAPI, 1 μg ml^{-1}). The stained filters were analysed manually; the total cell count includes heterotrophic bacteria as well as autofluorescent cyanobacteria, but not picoeukaryotic cells.

Concentration of Chl *a*

The concentration of Chl *a* was determined in subsurface water on each working day (Monday–Friday) as part of the Helgoland Roads LTER time series (https://www.awi.de/en/science/biosciences/shelf-sea-system-ecology/working-groups/long-term-observations-lto.html). The concentration of Chl *a* was assessed from fluorescence data using an algal group analyser (bbe moldaenke).

Determination of glycolate concentrations

Once per week, 5 aliquots of 2 ml each were taken from the filtrate after 0.2-μm filtration and stored at −80 °C until analysis. Glycolate concentrations were measured after derivatization of the samples with 3-nitrophenylhydrazine as previously described[72]. LC–MS analyses were performed on an Agilent 6495B Triple Quad LC–MS system equipped with an electrospray ionization source. The analytes were separated on a RP-18 column (50 mm × 2.1 mm, particle size 1.8 μm, ZORBAX RRHD Eclipse Plus C18; Agilent) kept at 40 °C using a mobile phase system that consisted of 0.1% formic acid in water (A) and acetonitrile (B). The gradient was as follows: 0 min, 5% B; 1 min, 5% B; 6 min, 95% B; 6.5 min, 95% B; 7 min, 5% B at a flow rate of 250 μl min^{-1}. Samples were held at 15 °C and injection volume was 5 μl. MS/MS data were acquired in negative MRM mode. Capillary voltage was set at 3 kV and nitrogen gas was used as nebulizing (25 psig), drying (11 l min^{-1}, 130 °C) and sheath gas (12 l min^{-1}, 400 °C). The dwell time and fragmentor voltage were 20 ms and 380 V, respectively. Optimized collision energy used for the derivatized glycolate (210 $m/z \rightarrow 137$ m/z) was 22 V. LC–MS data were analysed and quantified using MassHunter Qualitative Navigator and QQQ Quantitative Analysis software (Agilent).

Determination of glycolate uptake rates

Samples for glycolate uptake measurements were collected on 10 April, 15 May and 29 May 2018. All samples were used after filtration through a 3-μm filter and divided into 4 live 40 ml subsamples in sterile plastic tubes wrapped in aluminium foil and incubated with 165 nM calcium [1-^{14}C]glycolate (American Radiolabelled Chemicals; 55 mCi $mmol^{-1}$, 0.1 mCi ml^{-1} in sterile water) at 12 °C for 8 h. Controls consisted of four 40 ml subsamples killed in 10% formalin for 1 h before addition of 165 nM calcium [1-^{14}C]glycolate. Glycolate uptake was monitored over time by withdrawing 5 ml aliquots from each subsample, filtering each aliquot

onto a 0.2-µm pore size Nuclepore polycarbonate filter (GE Healthcare), rinsing the filter 3 times with 5 ml of filter-sterilized sea water and measuring the radioactivity with a Tri-Carb 4910 TR liquid scintillation analyser (PerkinElmer) using the Ultima Gold scintillation cocktail (PerkinElmer). Glycolate uptake rates were determined by linear fit of the counts per minute measured on the filters over time. Uptake rates were corrected to account for the presence of non-radioactive glycolate in the samples.

DNA and RNA extraction, cDNA synthesis and qPCR

DNA and RNA was extracted from filters using the AllPrep Bacterial DNA/RNA/Protein Kit (Qiagen) according to the manufacturer's instructions. The RNA samples were treated with the TURBO DNA-free Kit (Thermo Fisher Scientific) according to the manufacturer's instructions to exclude contamination with DNA. DNA and RNA concentrations were determined using the Qubit dsDNA/RNA HS Assay Kit (Thermo Fisher Scientific) according to the manufacturer's instructions. In total, 2 µg of RNA was used for cDNA synthesis with the GoScript Reverse Transcription System (Promega) and random hexamers according to the manufacturer's instructions.

Degenerate primers for the *bhcC* gene were designed using the j-CODEHOP software[73–75] and an alignment of 207 *bhcC* sequences from bacterial strains isolated from marine habitats. Sequences were aligned using MUSCLE. Extracted DNA and cDNA of RNA were quantified using a CFX Connect Real-Time System (Bio-Rad). SYBR Green JumpStart Taq ReadyMix (Sigma-Aldrich) was used for the PCR amplification mixture according to the manufacturer's instructions. Final $MgCl_2$ concentration was 3 mM, and the amplification protocol consisted of an initial enzyme activation step at 95 °C for 5 min, followed by 45 cycles of 95 °C for 30 s, 60 °C for 30 s, and 72 °C for 45 s. Eight standard amounts ranging from 3×10^1 to 3×10^8 copies were run in triplicate for each set of analyses. Regression of all standard curves yielded an r^2 value of at least 0.998. All samples were run in triplicate. The starting copy numbers of *bhcC* in DNA and cDNA were calculated based on regression parameters of standard curves, and gene/transcript copy numbers per cell were calculated based on the volume of sea water filtered, the microbial cell count at the time of sampling, the amount of extracted DNA or RNA, and the volume of DNA or cDNA used per reaction. The degenerate primers were validated with genomic DNA of *P. denitrificans* DSM 413, *Rhodobacter sphaeroides* 2.4.1, and *E. coli* K-12 MG1655 as template using the same qPCR protocol as above. Standards for quantification were created by PCR using genomic DNA of *P. denitrificans* DSM 413 as template. Purified *bhcC* PCR product was quantified using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) according to the manufacturer's instructions.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The coordinates and structure factors of the crystal structures generated from this research are available at the PDB under accession numbers 6QKB and 6RQA. Mass spectrometry proteomics data are available via ProteomeXchange with the identifier PXD013274. MAGs are available under accession PRJEB28156 at the European Nucleotide Archive (ENA). All other relevant data are available in the Article and the Supplementary Information. Source Data for Figs. 2, 3 and Extended Data Fig. 1, 4, 5, 7–9 are provided with the paper.

41. Bertani, G. Studies on lysogenesis. I. The mode of phage liberation by lysogenic *Escherichia coli*. *J. Bacteriol.* **62**, 293–300 (1951).
42. Beijerinck, M. W. & Minkman, D. C. J. Bildung und Verbrauch von Stickoxydul durch Bakterien. *Zentralbl. Bakteriol. Naturwiss.* **25**, 30–63 (1910).
43. Hahnke, S. M., Moosmann, P., Erb, T. J. & Strous, M. An improved medium for the anaerobic growth of *Paracoccus denitrificans* Pd1222. *Front. Microbiol.* **5**, 18 (2014).
44. Laemmli, U. K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680–685 (1970).
45. Kitagawa, M. et al. Complete set of ORF clones of *Escherichia coli* ASKA library (a complete set of *E. coli* K-12 ORF archive): unique resources for biological research. *DNA Res.* **12**, 291–299 (2005).
46. Lane, C. F. Sodium cyanoborohydride — a highly selective reducing agent for organic functional groups. *Synthesis* **1975**, 135–146 (1975).
47. Kiefer, P., Schmitt, U. & Vorholt, J. A. eMZed: an open source framework in Python for rapid and interactive development of LC/MS data analysis workflows. *Bioinformatics* **29**, 963–964 (2013).
48. Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72**, 248–254 (1976).
49. Ledermann, R., Strebel, S., Kampik, C. & Fischer, H. M. Versatile vectors for efficient mutagenesis of *Bradyrhizobium diazoefficiens* and other Alphaproteobacteria. *Appl. Environ. Microbiol.* **82**, 2791–2799 (2016).
50. Thoma, S. & Schobert, M. An improved *Escherichia coli* donor strain for diparental mating. *FEMS Microbiol. Lett.* **294**, 127–132 (2009).
51. Moggridge, S., Sorensen, P. H., Morin, G. B. & Hughes, C. S. Extending the compatibility of the SP3 paramagnetic bead processing approach for proteomics. *J. Proteome Res.* **17**, 1730–1740 (2018).
52. Glatter, T. et al. Large-scale quantitative assessment of different in-solution protein digestion protocols reveals superior cleavage efficiency of tandem Lys-C/trypsin proteolysis over trypsin digestion. *J. Proteome Res.* **11**, 5145–5156 (2012).
53. Engilberge, S. et al. Crystallophore: a versatile lanthanide complex for protein crystallography combining nucleating effects, phasing properties, and luminescence. *Chem. Sci.* **8**, 5909–5917 (2017).
54. Kabsch, W. XDS. *Acta Crystallogr. D* **66**, 125–132 (2010).
55. Winn, M. D. et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67**, 235–242 (2011).
56. Uhl, M. K. et al. The crystal structure of d-threonine aldolase from *Alcaligenes xylosoxidans* provides insight into a metal ion assisted PLP-dependent mechanism. *PLoS ONE* **10**, e0124056 (2015).
57. Gallagher, D. T. et al. Structure of alanine dehydrogenase from *Archaeoglobus*: active site analysis and relation to bacterial cyclodeaminases and mammalian mu crystallin. *J. Mol. Biol.* **342**, 119–130 (2004).
58. Waterhouse, A. et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).
59. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
60. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
61. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
62. Francis, T. B., Krüger, K., Fuchs, B. M., Teeling, H. & Amann, R. I. *Candidatus* Prosiliicoccus vernus, a spring phytoplankton bloom associated member of the *Flavobacteriaceae*. *Syst. Appl. Microbiol.* **42**, 41–53 (2019).
63. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
64. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
65. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
66. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
67. Silvestro, D. & Michalak, I. raxmlGUI: a graphical front-end for RAxML. *Org. Divers. Evol.* **12**, 335–337 (2012).
68. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
69. Gerlt, J. A. et al. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim. Biophys. Acta* **1854**, 1019–1037 (2015).
70. Villar, E. et al. The Ocean Gene Atlas: exploring the biogeography of plankton genes online. *Nucleic Acids Res.* **46**, W289–W295 (2018).
71. Kaplun, A. et al. Glyoxylate carboligase lacks the canonical active site glutamate of thiamine-dependent enzymes. *Nat. Chem. Biol.* **4**, 113–118 (2008).
72. Han, J., Gagnon, S., Eckle, T. & Borchers, C. H. Metabolomic analysis of key central carbon metabolism carboxylic acids as their 3-nitrophenylhydrazones by UPLC/ESI-MS. *Electrophoresis* **34**, 2891–2900 (2013).
73. Rose, T. M. et al. Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic Acids Res.* **26**, 1628–1635 (1998).
74. Rose, T. M., Henikoff, J. G. & Henikoff, S. CODEHOP (consensus-degenerate hybrid oligonucleotide primer) PCR primer design. *Nucleic Acids Res.* **31**, 3763–3766 (2003).
75. Boyce, R., Chilana, P. & Rose, T. M. iCODEHOP: a new interactive program for designing consensus-degenerate hybrid oligonucleotide primers from multiply aligned protein sequences. *Nucleic Acids Res.* **37**, W222–W228 (2009).

Acknowledgements We thank L. Franzmeyer and K.-P. Rücknagel as well as the crew of the *Aade* and the staff at the Biological Station at Helgoland (BAH) for sample collection and processing; K. H. Wiltshire for providing Chl *a* data collected at BAH; S. Vidal-Melgosa, A. Bolte and J.-H. Hehemann for sharing samples; T. Ferdelman for help with radioactive tracer work; B. Vögeli, T. Schwander, G. Stoffel and S. Burgener for helpful discussions. We acknowledge the support from the staff scientists at the European Synchrotron Radiation Facility Grenoble,

France (ESRF, beamlines ID29 & ID30B) as well as at the Deutsches Elektronen-Synchrotron Hamburg, Germany (DESY, beamline P13). Metagenome sequences were obtained within the COGITO (Coastal Microbe Genomic and Taxonomic Observatory) project granted to H. Teeling (MPI Bremen) as a community sequencing project by the Department of Energy's Joint Genome Institute in Walnut Creek, CA, USA (Proposal ID 998; https://doi.org/10.25585/1488076). This study was funded by the Max-Planck-Society (R.I.A. and T.J.E.), FET-Open Grant 686330 (Future Agriculture) and the German Research Foundation (SFB987 'Microbial diversity in environmental signal response' and FOR 2406 'Proteogenomics of marine polysaccharide utilization').

Author contributions L.S.v.B. identified the bhc gene cluster, purified proteins, performed enzyme kinetic analysis, qPCR, phylogenetic analysis and analysis of *Tara* Oceans metagenomes, generated and characterized mutant *P. denitrificans* strains and measured glycolate uptake rates. F. Severi performed enzyme kinetic analysis, crystallization of BhcD and enzyme assays in *P. denitrificans* cell-free extracts. K.K. performed phylogenetic analysis and analysis of Helgoland metagenomes. L.H. performed gel shift assays with BhcR. A.G. performed crystallization of BhcC. F. Sippel performed enzyme kinetic analysis. B.P. generated mutant *P. denitrificans* strains. P.C. and N.S.C. performed small-molecule mass spectrometry. T.G. performed mass spectrometry for proteomics. J.Z. collected X-ray datasets, solved, refined and analysed crystal structures. B.M.F. and R.I.A. planned and supervised fieldwork at Helgoland and provided reagents. L.S.v.B., E.B., S.Z., U.G.M., R.I.A. and T.J.E. planned experiments, analysed data and supervised the project. L.S.v.B. and T.J.E. wrote the manuscript, with contributions from all other authors.

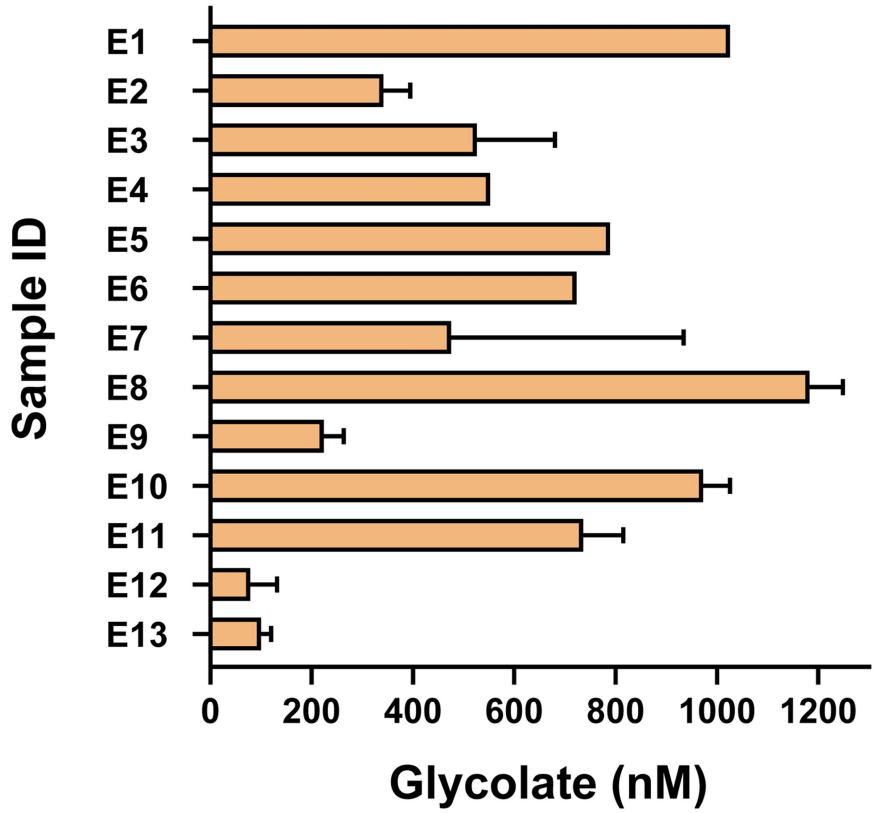
Competing interests The Max-Planck-Gesellschaft zur Förderung der Wissenschaften is the patent applicant for the following three patents. All patent applications are pending. L.S.v.B. and T.J.E. have filed European patent no. EP 19190404.4 for the production of plants with altered photorespiration due to implementation of the BHAC. L.S.v.B., J.Z. and T.J.E. have filed European patent no. EP 18167406.0 for the production of photoautotrophic organisms with altered photorespiration due to implementation of the BHAC. L.S.v.B. and T.J.E. have filed European patent no. 18211454.6 for the enantioselective preparation of primary amine compounds using the enzyme BhcD or its homologues.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1748-4.

Correspondence and requests for materials should be addressed to L.S.v.B. or T.J.E.

a

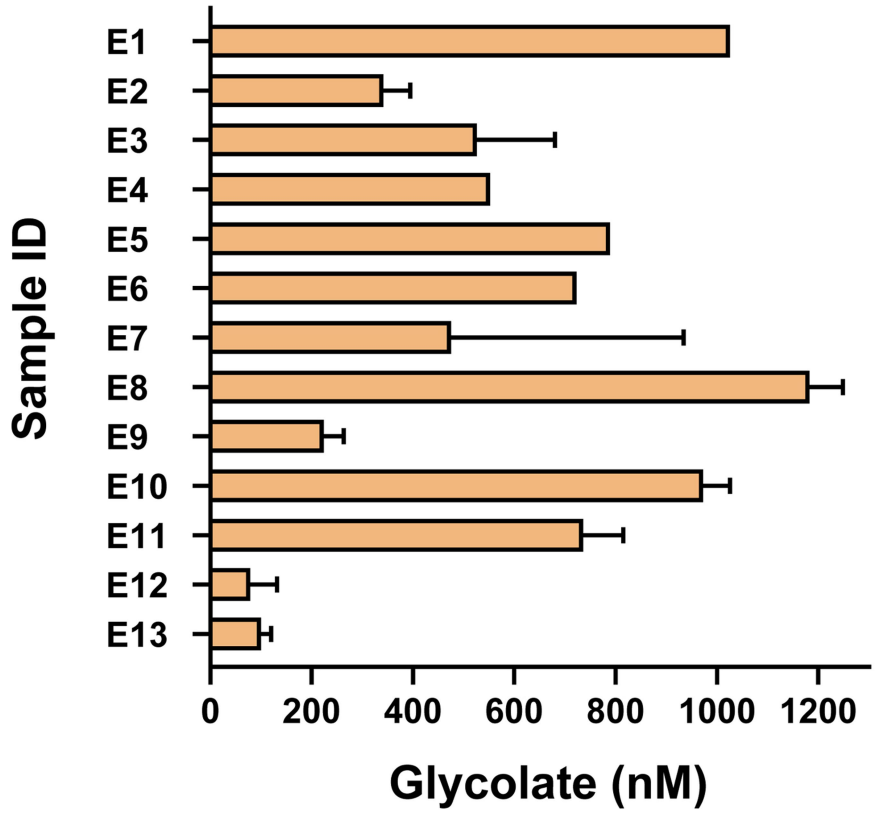


b

Sample ID	Sample	Maximum glycolate concentration [nM]	Analytical method	Reference
E1	Coastal seawater (Ipswich Bay, MA, USA)	1,026	Colorimetry	30
E2	Atlantic Ocean (51°40' S, 57°48' W)	342 ± 53	Colorimetry	31
E3	Antarctic lake (60°42' S, 45°37' W)	526 ± 154		
E4	Coastal seawater (Ipswich Bay, MA, USA)	552	Colorimetry	2
E5	Coastal seawater (Menai Straits, Anglesey, UK)	789	Colorimetry	32
E6	Coastal seawater (Ipswich Bay, MA, USA)	723	Colorimetry	1
E7	Coastal seawater (New York Bight, NY, USA)	475 ± 459	Colorimetry	33
E8	Mediterranean Sea (42°28' N, 30°16' E)	1,183 ± 66	HPLC	10
E9	Atlantic Ocean (oligotrophic waters, 21°01'54" N, 31°09'62" W)	224 ± 39	HPLC	11
E10	Atlantic Ocean (mesotrophic waters, 18°27'22" N, 21°10'18" W)	973 ± 53		
E11	Atlantic Ocean (eutrophic waters, 20°31'49" N, 18°34'39" W)	736 ± 79		
E12	Mediterranean Sea (43°25' N, 7°52' E)	79 ± 53	GC	34
E13	Coastal seawater (Dabob Bay, WA, USA)	100 ± 20	HPLC	4
C1	Culture of *Chlorella*	39,450 – 105,190	^{14}C-tracing	9
C2	Culture of *Euglena gracilis*	591,720	Colorimetry	35
C3	Culture of *Chaetoceros socialis*	240,630	Colorimetry	36
C4	Culture of *Dunaliella tertiolecta*	19,500 ± 1125	HPLC	37
C5	Culture of *Thalassiosira weissflogii*	799	HPLC	38
C6	Culture of *Prochlorococcus* MED4 (phosphorus-limited)	1873 ± 375	HPLC	39
C7	Culture of *Prochlorococcus* MED4 (phosphorus-replete)	2831 ± 458		
C8	Culture of *Prochlorococcus* MIT9312 (phosphorus-limited)	749 ± 916		
C9	Culture of *Prochlorococcus* MIT9312 (phosphorus-replete)	333 ± 125		

Extended Data Fig. 1 | Previously reported glycolate concentrations in environmental samples and cultures of photosynthetic organisms. a, Bar diagram of glycolate concentrations as previously reported in environmental samples. For details on samples, replicates, and analytics see **b** and the literature cited therein. **b**, Table of glycolate concentrations as previously reported in environmental samples (E1, E2 and so on) and cultures of photosynthetic organisms (C1, C2 and so on). When reported in the reference[1,2,4,9–11,30–39], the mean value ± error is given.

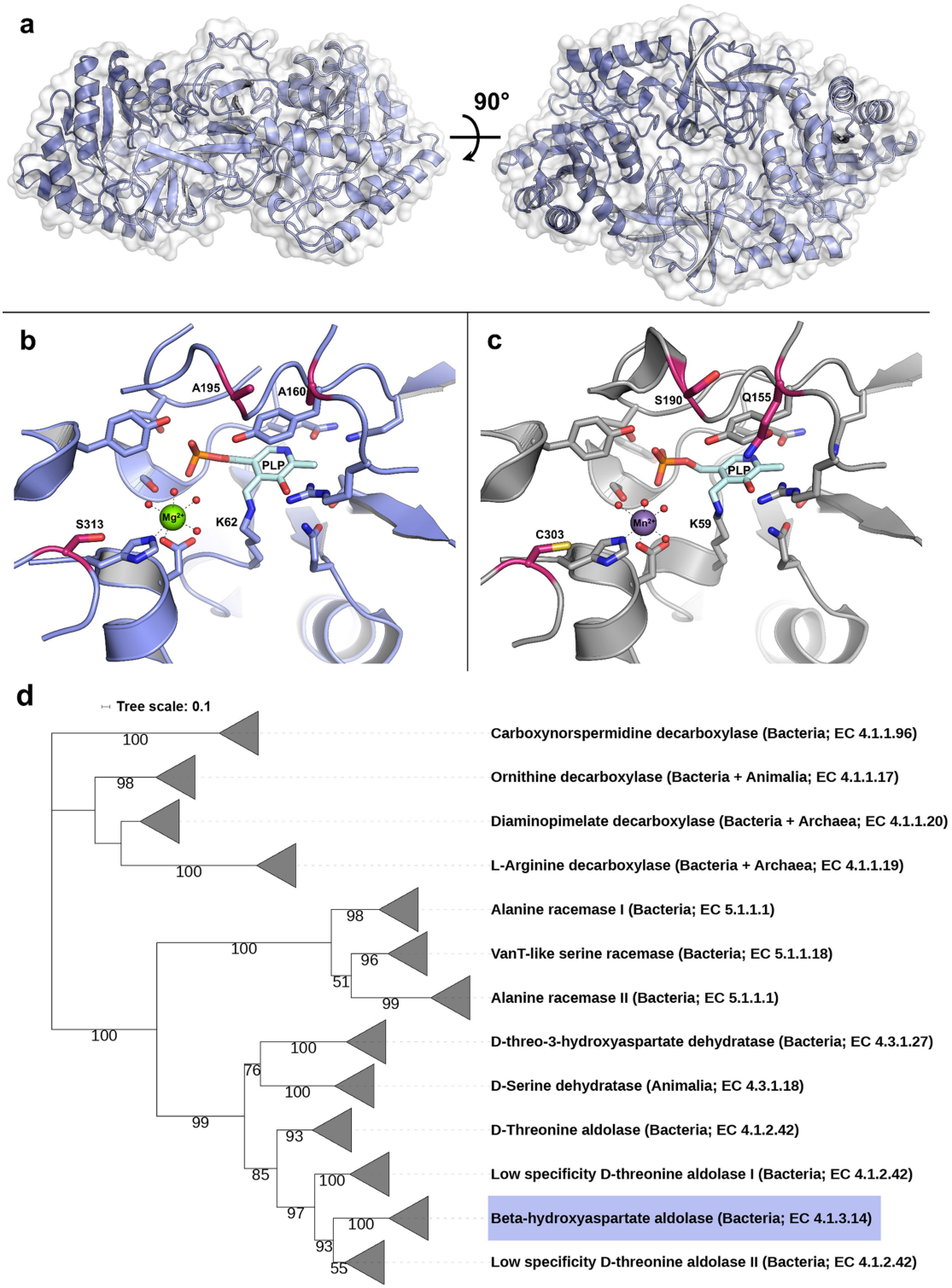
a



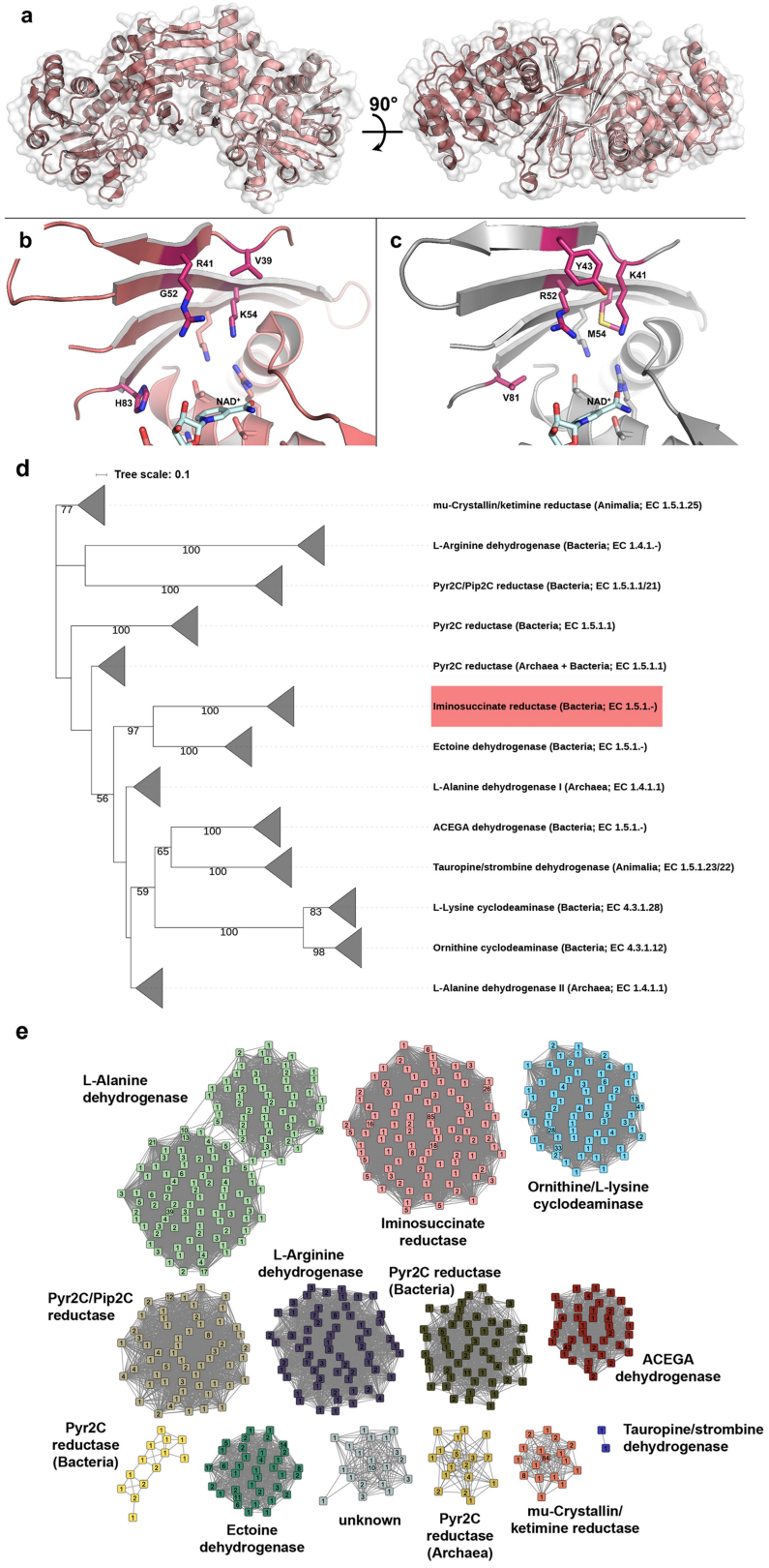
b

Sample ID	Sample	Maximum glycolate concentration [nM]	Analytical method	Reference
E1	Coastal seawater (Ipswich Bay, MA, USA)	1,026	Colorimetry	30
E2	Atlantic Ocean (51°40' S, 57°48' W)	342 ± 53	Colorimetry	31
E3	Antarctic lake (60°42' S, 45°37' W)	526 ± 154		
E4	Coastal seawater (Ipswich Bay, MA, USA)	552	Colorimetry	2
E5	Coastal seawater (Menai Straits, Anglesey, UK)	789	Colorimetry	32
E6	Coastal seawater (Ipswich Bay, MA, USA)	723	Colorimetry	1
E7	Coastal seawater (New York Bight, NY, USA)	475 ± 459	Colorimetry	33
E8	Mediterranean Sea (42°28' N, 30°16' E)	1,183 ± 66	HPLC	10
E9	Atlantic Ocean (oligotrophic waters, 21°01'54" N, 31°09'62" W)	224 ± 39	HPLC	11
E10	Atlantic Ocean (mesotrophic waters, 18°27'22" N, 21°10'18" W)	973 ± 53		
E11	Atlantic Ocean (eutrophic waters, 20°31'49" N, 18°34'39" W)	736 ± 79		
E12	Mediterranean Sea (43°25' N, 7°52' E)	79 ± 53	GC	34
E13	Coastal seawater (Dabob Bay, WA, USA)	100 ± 20	HPLC	4
C1	Culture of *Chlorella*	39,450 – 105,190	^{14}C-tracing	9
C2	Culture of *Euglena gracilis*	591,720	Colorimetry	35
C3	Culture of *Chaetoceros socialis*	240,630	Colorimetry	36
C4	Culture of *Dunaliella tertiolecta*	19,500 ± 1125	HPLC	37
C5	Culture of *Thalassiosira weissflogii*	799	HPLC	38
C6	Culture of *Prochlorococcus* MED4 (phosphorus-limited)	1873 ± 375	HPLC	39
C7	Culture of *Prochlorococcus* MED4 (phosphorus-replete)	2831 ± 458		
C8	Culture of *Prochlorococcus* MIT9312 (phosphorus-limited)	749 ± 916		
C9	Culture of *Prochlorococcus* MIT9312 (phosphorus-replete)	333 ± 125		

Extended Data Fig. 1 | Previously reported glycolate concentrations in environmental samples and cultures of photosynthetic organisms. a, Bar diagram of glycolate concentrations as previously reported in environmental samples. For details on samples, replicates, and analytics see **b** and the literature cited therein. **b**, Table of glycolate concentrations as previously reported in environmental samples (E1, E2 and so on) and cultures of photosynthetic organisms (C1, C2 and so on). When reported in the reference[1,2,4,9–11,30–39], the mean value ± error is given.

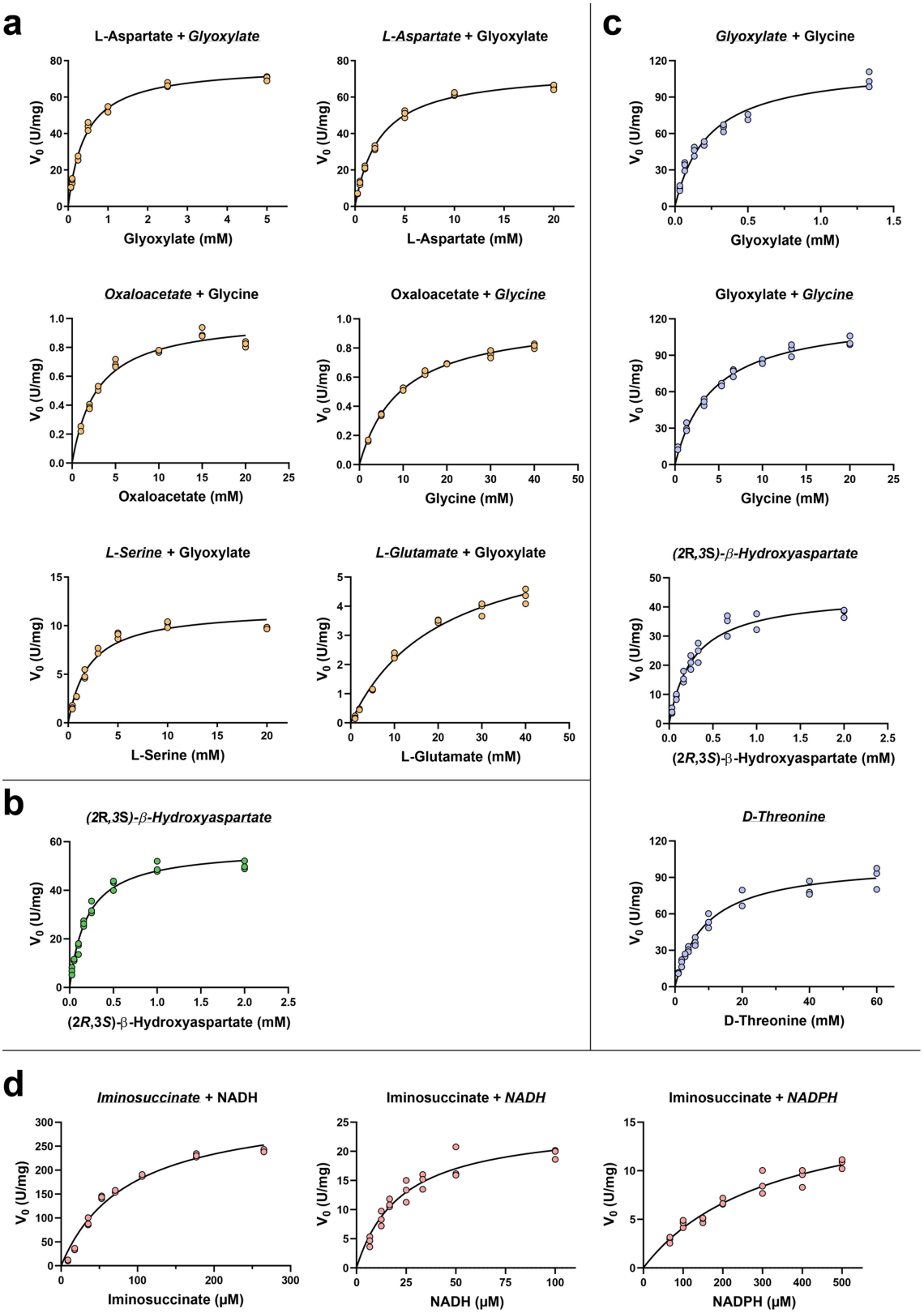


Extended Data Fig. 2 | Crystal structure and phylogenetic analysis of the β-hydroxyaspartate aldolase BhcC. a, Cartoon representation of the β-hydroxyaspartate aldolase homodimer (PDB 6QKB) with superimposed protein surface (left, side view; right, top view). **b**, Active site of β-hydroxyaspartate aldolase with covalently bound PLP (light cyan). Active site residues highlighted in pink (A160, A195 and S313) are completely conserved only among β-hydroxyaspartate aldolases, but differ in D-threonine aldolases. **c**, Active site of D-threonine aldolase (PDB 4V15). The corresponding conserved residues among D-threonine aldolases (Q155, S190 and C303) are highlighted as in **b**. **d**, Maximum likelihood phylogenetic tree of the type III PLP-dependent protein superfamily. Sequences of the β-hydroxyaspartate aldolase BhcC and its homologues form a distinct clade (blue) within the D-threonine aldolase branch of this superfamily. Bootstrap values of at least 50 are given on the respective nodes.

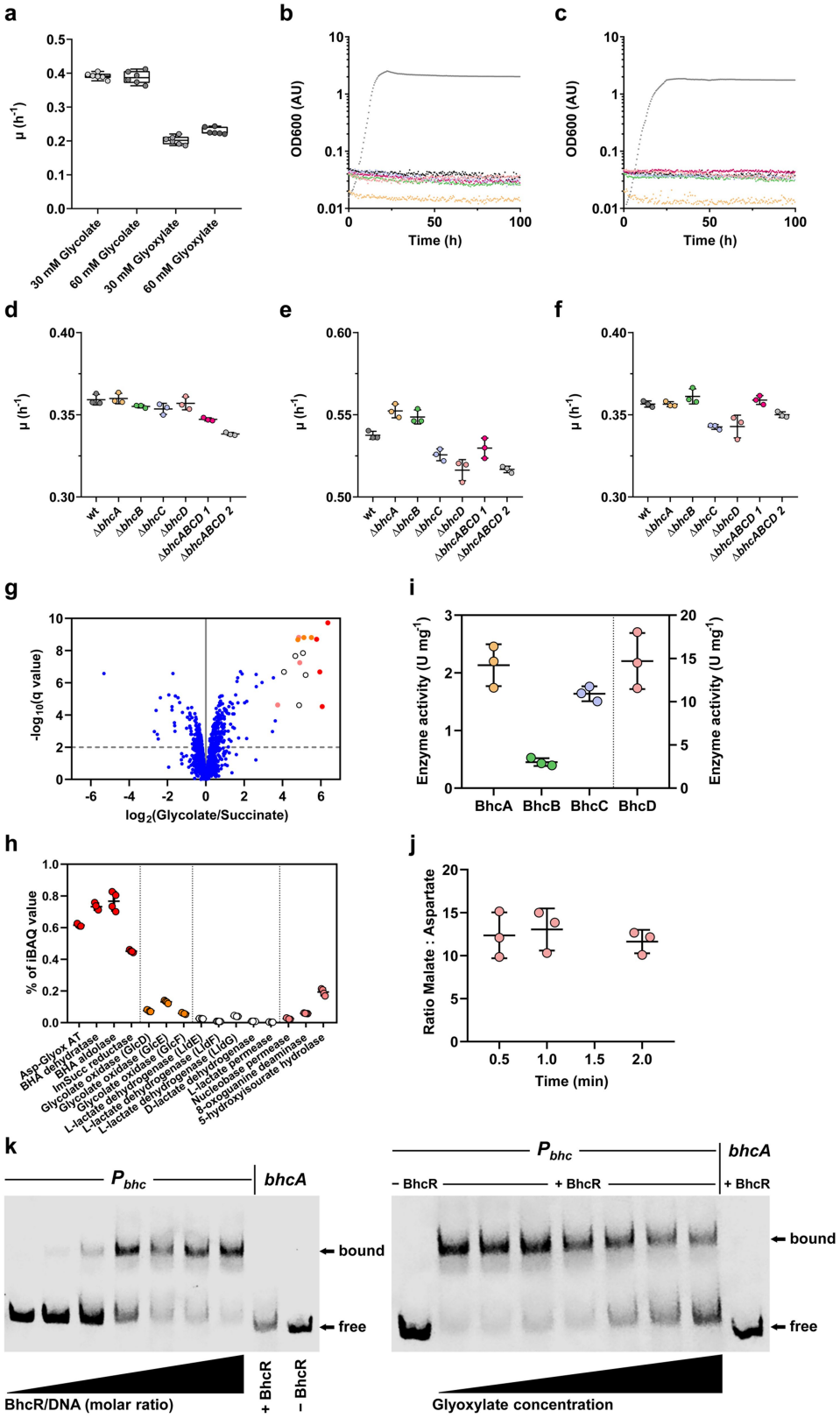


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Crystal structure and phylogenetic analysis of the iminosuccinate reductase BhcD. a, Cartoon representation of the iminosuccinate reductase homodimer (PDB 6RQA) with superimposed protein surface (left, side view; right, top view). **b**, Active site of BhcD with bound NAD^+ (light cyan). Residues highlighted in pink (V39, R41, G52, K54 and H83) may contribute to substrate binding and are conserved among iminosuccinate reductases, but differ in L-alanine dehydrogenases. **c**, Active site of L-alanine dehydrogenase (PDB 1OMO). The corresponding conserved residues among L-alanine dehydrogenases (K41, Y43, R52, M54 and V81) are highlighted as in **b**. **d**, Maximum likelihood phylogenetic tree of the ornithine cyclodeaminase/μ-crystalline protein superfamily. Sequences of the iminosuccinate reductase BhcD and its homologues form a distinct clade (red) within this superfamily. Bootstrap values of at least 50 are given on the respective nodes. **e**, Sequence similarity network of 1,614 sequences from the ornithine cyclodeaminase/μ-crystalline protein superfamily. Connected sequences with more than 80% identity are clustered into nodes. The number in each node gives the number of sequences contained within. Nodes with more than 50% identity are connected by edges. Similar to the phylogenetic analysis shown in **d**, sequences of the iminosuccinate reductase BhcD and its homologues form a distinct clade (red) within this superfamily.

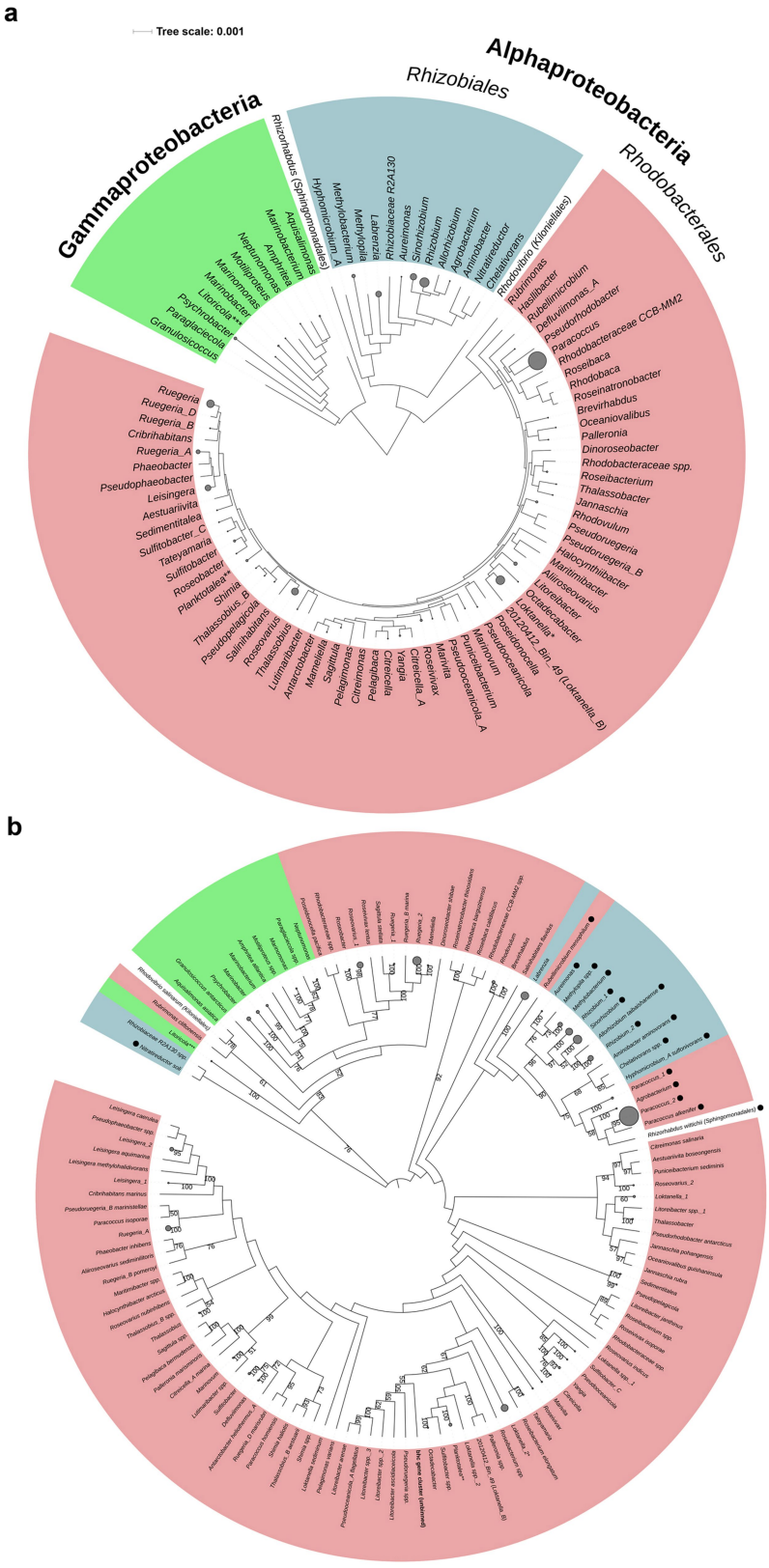


Extended Data Fig. 4 | Michaelis–Menten kinetics of all enzyme reactions characterized in this study. a, Michaelis–Menten kinetics for aspartate–glyoxylate aminotransferase (BhcA). **b**, Michaelis–Menten kinetics for β-hydroxyaspartate dehydratase (BhcB). **c**, Michaelis–Menten kinetics for β-hydroxyaspartate aldolase (BhcC). **d**, Michaelis–Menten kinetics for iminosuccinate reductase (BhcD). **a**–**d**, Data are shown from $n = 3$ independent experiments at different substrate concentrations. The data are summarized in Table 1.



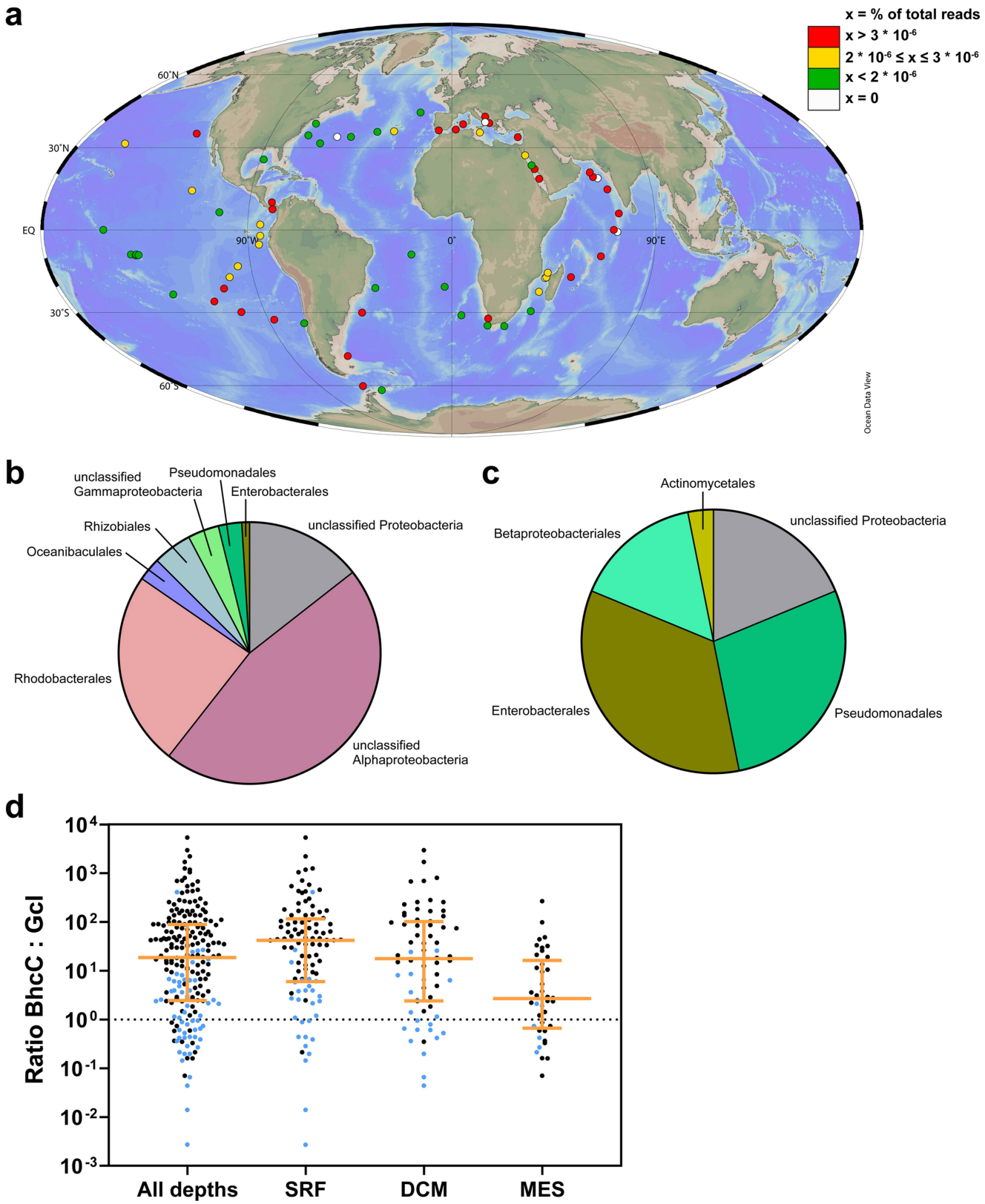
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Physiological role of the BHAC in *P. denitrificans* DSM 413. a, Growth rate of wild-type *P. denitrificans* DSM 413 on the BHAC substrates glycolate and glyoxylate. The middle line and box are the median and interquartile range of $n = 6$ independent experiments and the whiskers indicate the maximum range of the dataset. **b**, **c**, Representative growth curves of wild-type *P. denitrificans* DSM 413 (grey) and bhc deletion strains (coloured) grown in the presence of 60 mM glycolate (**b**) or 60 mM glyoxylate (**c**). Deletion of any single gene in the bhc gene cluster is sufficient to completely abolish growth in the presence of glycolate and glyoxylate. These experiments were repeated three times independently with similar results. **d**–**f**, Growth rates (μ) of wild-type *P. denitrificans* DSM 413 (grey) and BHAC deletion strains (coloured) grown in the presence of 60 mM acetate (**d**), 30 mM succinate (**e**) or 20 mM glucose (**f**). Deletion of any single gene in the bhc gene cluster, or of the whole bhc gene cluster, still permits growth on acetate, succinate or glucose with comparable growth rates as for the wild type. Data are the mean ± s.d. of $n = 3$ independently grown cultures. **g**, Analysis of the proteome of glycolate-grown compared to succinate-grown *P. denitrificans* DSM 413. All proteins that were quantified by at least three unique peptides are shown. The 15 proteins that showed the strongest increase in abundance are marked in the volcano plot. The four enzymes of the BHAC are marked in red, the three subunits of glycolate oxidase in orange, the proteins of a putative operon for lactate utilization in white and the proteins directly downstream of the bhc gene cluster in light red. **h**, The abundance of these proteins, given as the percentage of the intensity-based absolute quantification (iBAQ) value. Data are the mean ± s.d. of $n = 4$ independently grown cultures. **i**, Specific activities of BHAC enzymes in cell-free extracts of glycolate-grown *P. denitrificans* DSM 413, as measured spectrophotometrically. Note that the activity of BhcD is plotted on the right *y* axis and consists of the actual iminosuccinate reductase activity (iminosuccinate to L-aspartate) as well as endogenous malate dehydrogenase activity (oxaloacetate to L-malate). **j**, Ratio of malate to aspartate determined by LC–MS during the enzyme assay for BhcD activity. The ratio remains approximately constant at 12:1, indicating that only approximately 8% of the activity (around 1.3 U mg^{-1}) shown in **i** can be ascribed to iminosuccinate reductase. **i**, **j**, Data are the mean ± s.d. of $n = 3$ independently grown cultures; each data point represents the mean of $n = 3$ technical replicates. **k**, DNA-binding properties of BhcR. Left, a fluorescently labelled DNA fragment carrying the putative promoter region of the bhc gene cluster (P_{bhc}) was incubated with increasing amounts of purified BhcR protein and subsequently separated by electrophoresis to visualize DNA bound to BhcR and free DNA; a DNA fragment derived from the coding region of *bhcA* was used as a negative control. BhcR specifically forms a complex with the DNA fragment containing the putative promoter region of the bhc gene cluster. Right, the P_{bhc}–BhcR complex was incubated with increasing concentrations of glyoxylate and subsequently separated by electrophoresis to assess the effect of glyoxylate on complex formation; the *bhcA* DNA fragment together with BhcR was used as a negative control. Increasing concentrations of glyoxylate decrease the binding of BhcR to the P_{bhc} DNA fragment. For gel source data, see Supplementary Fig. 1.



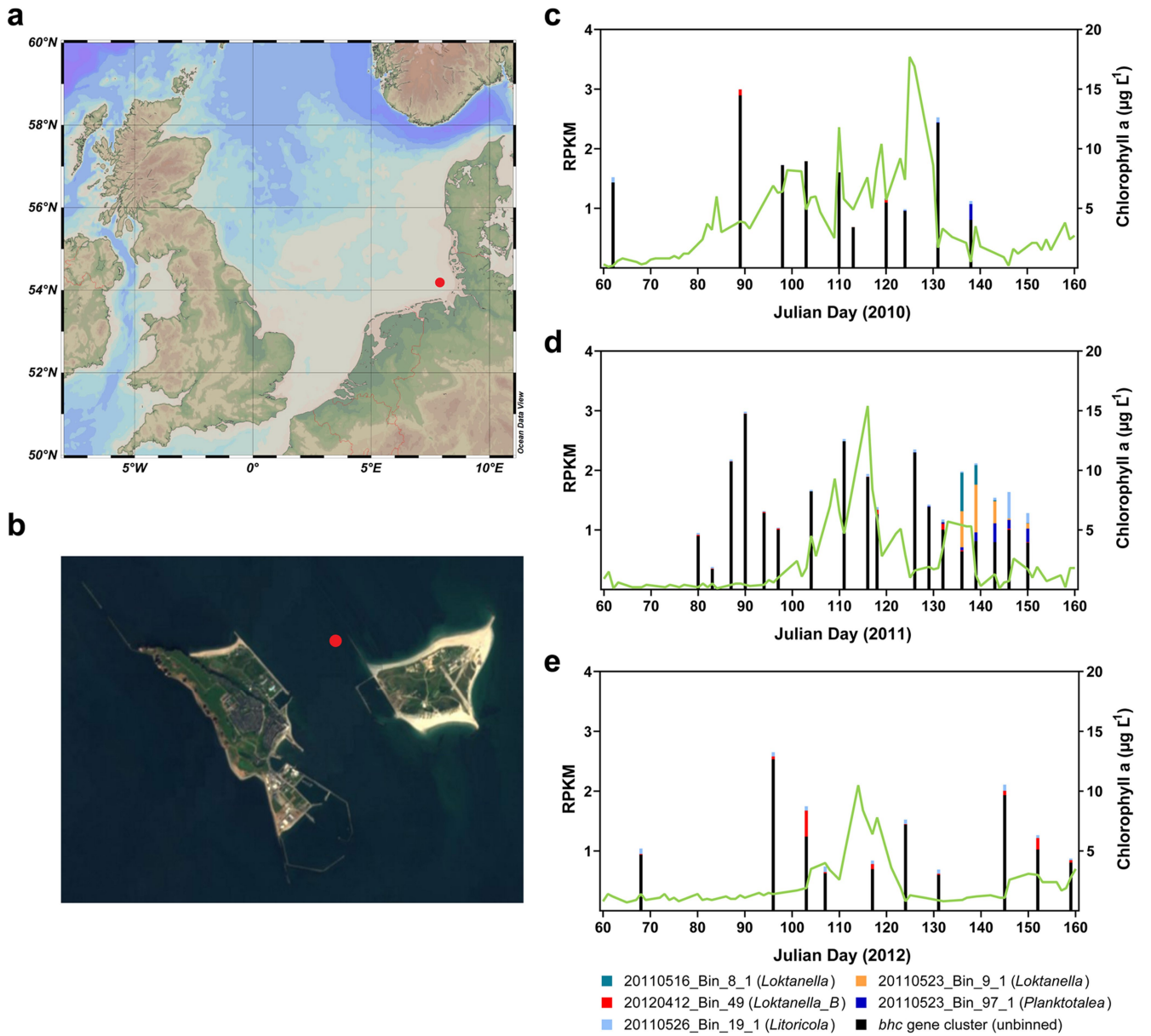
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Phylogenetic analysis of the bhc gene cluster. **a**, Genome-based maximum likelihood phylogenetic tree of bacterial strains with the bhc gene cluster. The bhc gene cluster is found in Gammaproteobacteria (green), and in the alphaproteobacterial orders Rhizobiales (blue) and Rhodobacterales (red), as well as in one member each of Sphingomonadales and Kiloniellales. The phylogenetic tree is based on an alignment of 120 bacterial marker genes from 264 publicly available bacterial genomes and 5 MAGs and was calculated using GTDB-Tk[64] (https://github.com/Ecogenomics/GtdbTk). If several strains from the same genus cluster together, nodes are collapsed at the genus level, and the size of the resulting circle corresponds to the respective number of strains. Loktanella*: collapsed node contains the MAGs 20110516_Bin_8_1 and 20110523_Bin_9_1; Planktotalea**: collapsed node contains the MAG 20110523_Bin_97_1; Litoricola***: collapsed node contains the MAG 20110526_Bin_19_1. **b**, Maximum likelihood phylogenetic tree of concatenated BHAC enzyme sequences. Colour code is the same as in **a**. Phylogenetic groups that were mostly isolated from terrestrial or freshwater habitats are marked with a black dot. Comparison with **a** reveals that the sequences of the BHAC enzymes are not phylogenetically representative, as, for example, alpha- and gammaproteobacterial sequences form a common branch and sequences from terrestrial or freshwater Rhizobiales and Rhodobacterales form another common branch. This suggests that the bhc gene cluster might have been subject to horizontal gene transfer between distantly related strains in shared habitats. The environmental bhc gene cluster sequence that could not be binned successfully is marked in bold and clusters together with isolated representatives of *Pseudoruegeria*, *Litoreibacter* and *Pseudooceanicola*. The phylogenetic tree is based on the concatenated alignments of the 4 enzymes (BhcA–BhcD) from 264 publicly available bacterial genomes and from 6 metagenome contigs. It was calculated using raxmlGUI[67]. Bootstrap values of at least 50 are given on the respective nodes; calculated branch lengths of the tree are ignored for the sake of better visualization. If several strains from the same genus cluster together, nodes are collapsed at the genus level, and the size of the resulting circle corresponds to the respective number of strains. If strains from the same genus cluster in more than one node, the respective branches are labelled as Genus_1, Genus_2, and so on, in a clockwise manner. Loktanella_2*: collapsed node contains the MAGs 20110516_Bin_8_1 and 20110523_Bin_9_1; Planktotalea**: collapsed node contains the MAG 20110523_Bin_97_1; Litoricola***: collapsed node contains the MAG 20110526_Bin_19_1. **a**, **b**, Taxonomy is based on GTDB (release 03-RS86; http://gtdb.ecogenomic.org/). All strains contained in the phylogenetic trees are listed in Supplementary Data 1.

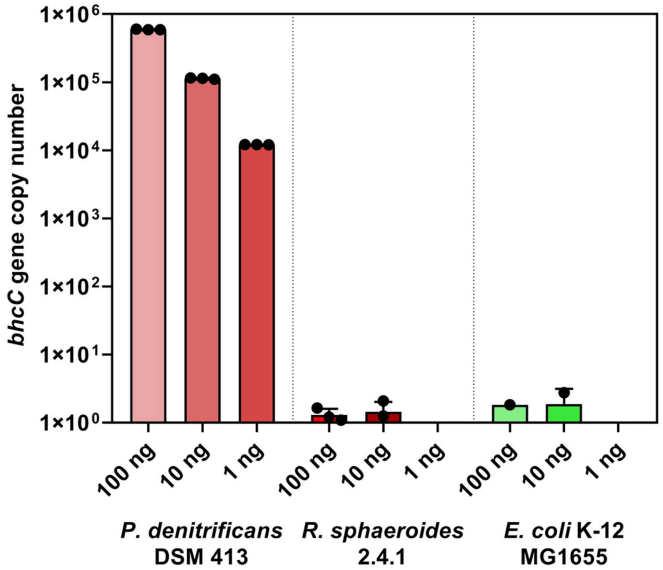


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Glyoxylate assimilation pathways in marine metagenomes. a, Metagenomes collected during the *Tara* Oceans expedition were searched for the presence of BhcC as representative enzyme of the BHAC. Dots on the map mark sampling locations of metagenomes containing BhcC sequences; the colour of the dot corresponds to BhcC abundance in samples from surface water (0.22–3-µm size fraction), as shown in the legend. The map was made with Ocean Data View 5.1.5 (Schlitzer, R., Ocean Data View, odv.awi.de, 2018). **b**, Phylogenetic distribution of 104 unique BhcC sequences found in *Tara* Oceans metagenomes. **c**, Phylogenetic distribution of 32 unique Gcl (as representative enzyme of the glycerate pathway) sequences found in *Tara* Oceans metagenomes. Whereas BhcC is mainly found in Alphaproteobacteria, Gcl is largely restricted to Gammaproteobacteria. **b**, **c**, Taxonomy is based on GTDB[64] (release 03-RS86; http://gtdb.ecogenomic.org/). **d**, Ratio of the abundances (in percentage of total reads) of BhcC to Gcl in *Tara* Oceans metagenomes. BhcC:Gcl ratios from $n = 210$ samples are plotted together (left) and clustered by sampling depth (SRF, upper layer zone ($n = 101$); DCM, deep chlorophyll maximum layer ($n = 68$); MES, mesopelagic zone ($n = 41$)). Samples from the 0.22–3-µm size fraction are denoted by a black dot; samples from the <0.22-µm size fraction are denoted by a blue dot. The median is shown in orange as centre value and error bars represent interquartile ranges. The median BhcC:Gcl ratio of all samples is 18.7. The highest BhcC:Gcl ratio is found in surface water samples (median = 41.8), with the ratio generally being higher in the 0.22–3-µm size fraction than in the <0.22-µm size fraction. Sequence IDs, abundances and BhcC:Gcl ratios are given in Supplementary Data 2.



Extended Data Fig. 8 | Abundance of the bhc gene cluster in Helgoland metagenomes. a, The location of Helgoland Island approximately 40 km offshore the northern German coastline in the North Sea is marked with a red dot. The map was made with Ocean Data View 5.1.5 (R. Schlitzer, Ocean Data View, odv.awi.de, 2018). **b**, The long-term ecological research site 'Kabeltonne' (red dot: 54° 11.3′ N, 7° 54.0′ E) is located between Helgoland Island (left) and the small island Düne (right). Satellite image from WorldWind Explorer (B. Schubert, worldwind.earth/explorer, 2016–2018); the image was adapted to indicate the sampling site. **c**–**e**, Abundance of the bhc gene cluster (in RPKM) was calculated in 38 metagenomes from samples collected during the algal spring blooms of 2010 to 2012 in the North Sea close to Helgoland[28]. Six different sequences were investigated, five of which could be assigned to metagenome bins (Extended Data Fig. 6 and Supplementary Data 3), whereas the remaining, most abundant sequence (black) could not be binned successfully.



Extended Data Fig. 9 | Validation of degenerate *bhcC* primers. Degenerate primers for *bhcC* were used for qPCR with different amounts of genomic DNA from *P. denitrificans* DSM 413, *Rhodobacter sphaeroides* 2.4.1, and *E. coli* K-12 MG1655 as template. While the *bhcC* gene from *P. denitrificans* DSM 413 is amplified, genomic DNA from organisms that lack the bhc gene cluster does not result in reliable amplification. Data are mean ± s.d.; $n = 3$ independent experiments.

Extended Data Table 1 | X-ray diffraction data collection and model refinement statistics

	β-hydroxyaspartate aldolase with bound pyridoxalphosphate (PDB ID 6QKB)	iminosuccinate reductase with bound NAD^+ (PDB ID 6RQA)
Data collection		
Space group	$P\,2_1\,2_1\,2_1$	$P\,2_1\,2_1\,2_1$
Cell dimensions		
a, *b*, *c* (Å)	66.60, 75.25, 157.31	50.39, 72.41, 164.27
α, β, γ (°)	90.00, 90.00, 90.00	90.00, 90.00, 90.00
Resolution (Å)	29.03 - 1.70 (1.79 - 1.70)	29.40 – 2.56 (2.70 – 2.56)
R_{merge}	0.134 (0.858)	0.097 (0.527)
$I / \sigma I$	10.4 (1.9)	12.6 (3.3)
$CC_{1/2}$ (%)	99.7 (70.8)	99.8 (90.7)
Completeness (%)	99.8 (99.0)	99.9 (100.0)
Redundancy	6.7 (6.5)	6.5 (6.5)
Refinement		
Resolution (Å)	29.03 - 1.70 (1.74 - 1.70)	29.40 – 2.56 (2.70 – 2.56)
No. unique reflections	87194 (5909)	20152 (2893)
R_{work} / R_{free}	0.158 / 0.177	0.176 / 0.225
No. atoms	6671	5027
Protein	5817	4780
Ligands	32	120
Water	822	127
B-factors		
Protein	17.05	52.02
Ligands	23.66	66.03
Water	31.58	47.34
R.m.s. deviations		
Bond lengths (Å)	0.006	0.004
Bond angles (°)	0.84	0.52

Numbers in parentheses indicate statistics for the highest resolution shell. The structures were determined from single crystals.

nature research

Corresponding author(s): Lennart Schada von Borzyskowski, Tobias J. Erb

Last updated by author(s): Sep 4, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data collection and analysis was performed using publicly available software as detailed in citations included in the manuscript and SI. Cary WinUV 5.0.0.999; GraphPad Prism 8.0.0; Excel 2013; BBMap 35.14; GTDBtk 0.1.3; raxmlGUI 1.5b2; XDS BUILT 20180126 ; CCP4 7.0; Phenix 1.14_3260; Coot 0.8.9; Pymol 1.8; eMZed 2.29.4.0; Progenesis QI 2.0; MASCOT 2.5; SafeQuant 2.2.2; iTOL 4.3.2; MUSCLE 3.8.31; Ocean Gene Atlas (tara-oceans.mio.osupytheas.fr/ocean-gene-atlas); Ocean Data View 5.1.5; MassHunter B.09.00; Cytoscape 3.7.1; SWISS-MODEL (swissmodel.expasy.org); CheckM 1.0.7; EFI-EST (efi.igb.illinois.edu/efi-est/index.php); QQQ Quantitative Analysis; j-CODEHOP (4virology.net/virology-ca-tools/j-codehop)
Data analysis	Data collection and analysis was performed using publicly available software as detailed in citations included in the manuscript and SI. Cary WinUV 5.0.0.999; GraphPad Prism 8.0.0; Excel 2013; BBMap 35.14; GTDBtk 0.1.3; raxmlGUI 1.5b2; XDS BUILT 20180126 ; CCP4 7.0; Phenix 1.14_3260; Coot 0.8.9; Pymol 1.8; eMZed 2.29.4.0; Progenesis QI 2.0; MASCOT 2.5; SafeQuant 2.2.2; iTOL 4.3.2; MUSCLE 3.8.31; Ocean Gene Atlas (tara-oceans.mio.osupytheas.fr/ocean-gene-atlas); Ocean Data View 5.1.5; MassHunter B.09.00; Cytoscape 3.7.1; SWISS-MODEL (swissmodel.expasy.org); CheckM 1.0.7; EFI-EST (efi.igb.illinois.edu/efi-est/index.php); QQQ Quantitative Analysis; j-CODEHOP (4virology.net/virology-ca-tools/j-codehop)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The coordinates and structure factors of the crystal structures generated from this research (Fig. 1, Extended Data Fig. 2 + 3) are available at the Protein Data Bank

under accession numbers 6QKB and 6RQA. Mass spectrometry proteomics data (Extended Data Fig. 5) are available via ProteomeXchange with the identifier PXD013274. Metagenome-assembled genomes (MAGs; Extended Data Fig. 6 + 8) are available under accession PRJEB28156 at the European Nucleotide Archive (ENA). All other relevant data are available in this article and its Supplementary Information files.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. The common rationale was to have a sample size that allows us to calculate a standard deviation. All Michaelis-Menten plots for the kinetic characterization of enzymes include 18 data points (three independent experiments at 6 different substrate concentrations). Controls verifying same levels of specific activities between different enzyme preparations were performed routinely. Three independent experiments (= assays in independent cuvettes) were conducted for the determination of enzyme activities in vitro, both with purified enzymes and in P. denitrificans cell extracts. Three biological replicates (= independent cultures) of P. denitrificans were used to generate cell-free extracts for determination of enzyme activities. Five biological replicates (= independent water samples) were used to determine glycolate concentrations in seawater. Three technical replicates (= repeated injections of the same sample) were measured for the determination of glycolate concentrations in each single seawater sample via LC-MS. Three independent experiments (= independent wells of 96-well qPCR plate) were conducted for all samples and standards in qPCR experiments. Three or six biological replicates (= independent cultures in different wells of 96 well-plate) were measured for the determination of P. denitrificans growth rates. Four biological replicates (= independent cultures) per condition were used to generate biomass for proteomic analysis. Four biological replicates (= independent water samples) were used to determine glycolate uptake rates.
Data exclusions	No data were excluded from the analyses.
Replication	Controls verifying same levels of specific activities between different enzyme preparations were performed routinely. When applicable, we performed our experiments using multiple independent samples (e.g., independent cultures, independent water samples). All attempts at replication of our findings were successful.
Randomization	The microorganisms used in this study were selected and divided randomly in the different conditions. No criteria of selection were applied.
Blinding	Blinding of samples was not applicable for the kind of experiments included in this study. Blinding was not possible because the person in charge of the analysis was the one responsible for taking the samples.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

nature research

Corresponding author(s): Lennart Schada von Borzyskowski, Tobias J. Erb

Last updated by author(s): Sep 4, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data collection and analysis was performed using publicly available software as detailed in citations included in the manuscript and SI. Cary WinUV 5.0.0.999; GraphPad Prism 8.0.0; Excel 2013; BBMap 35.14; GTDBtk 0.1.3; raxmlGUI 1.5b2; XDS BUILT 20180126 ; CCP4 7.0; Phenix 1.14_3260; Coot 0.8.9; Pymol 1.8; eMZed 2.29.4.0; Progenesis QI 2.0; MASCOT 2.5; SafeQuant 2.2.2; iTOL 4.3.2; MUSCLE 3.8.31; Ocean Gene Atlas (tara-oceans.mio.osupytheas.fr/ocean-gene-atlas); Ocean Data View 5.1.5; MassHunter B.09.00; Cytoscape 3.7.1; SWISS-MODEL (swissmodel.expasy.org); CheckM 1.0.7; EFI-EST (efi.igb.illinois.edu/efi-est/index.php); QQQ Quantitative Analysis; j-CODEHOP (4virology.net/virology-ca-tools/j-codehop)
Data analysis	Data collection and analysis was performed using publicly available software as detailed in citations included in the manuscript and SI. Cary WinUV 5.0.0.999; GraphPad Prism 8.0.0; Excel 2013; BBMap 35.14; GTDBtk 0.1.3; raxmlGUI 1.5b2; XDS BUILT 20180126 ; CCP4 7.0; Phenix 1.14_3260; Coot 0.8.9; Pymol 1.8; eMZed 2.29.4.0; Progenesis QI 2.0; MASCOT 2.5; SafeQuant 2.2.2; iTOL 4.3.2; MUSCLE 3.8.31; Ocean Gene Atlas (tara-oceans.mio.osupytheas.fr/ocean-gene-atlas); Ocean Data View 5.1.5; MassHunter B.09.00; Cytoscape 3.7.1; SWISS-MODEL (swissmodel.expasy.org); CheckM 1.0.7; EFI-EST (efi.igb.illinois.edu/efi-est/index.php); QQQ Quantitative Analysis; j-CODEHOP (4virology.net/virology-ca-tools/j-codehop)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The coordinates and structure factors of the crystal structures generated from this research (Fig. 1, Extended Data Fig. 2 + 3) are available at the Protein Data Bank

under accession numbers 6QKB and 6RQA. Mass spectrometry proteomics data (Extended Data Fig. 5) are available via ProteomeXchange with the identifier PXD013274. Metagenome-assembled genomes (MAGs; Extended Data Fig. 6 + 8) are available under accession PRJEB28156 at the European Nucleotide Archive (ENA). All other relevant data are available in this article and its Supplementary Information files.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Investigation of spring phytoplankton bloom 2018 at Helgoland island. Data on chlorophyll A concentrations, total cell count, glycolate concentrations and uptake rates, and abundance of bhcC genes and transcripts in microbial biomass from water samples are quantitative.
Research sample	Water samples were taken from a designated location (see below) close to Helgoland island; no statistical method was used to predetermine sample size. This sampling site was chosen because it had been used in previous sampling campaigns by the MPI Bremen and the Biologische Anstalt Helgoland. Therefore, it was relevant to take water samples in 2018 that could be compared to samples taken at the same site in previous years.
Sampling strategy	Five biological replicates (= independent water samples) were used to determine glycolate concentrations in seawater. Four biological replicates (= independent water samples) were used to determine glycolate uptake rates. One water sample was used to determine total cell count and chlorophyll A concentration. One water sample was used to obtain microbial biomass for DNA and RNA extraction. No sample size calculation was performed. The common rationale for replicate samples was to have a sample size that allows us to calculate a standard deviation.
Data collection	Sampling was carried out on each working day (Monday - Friday) by the crew of the RV Aade (www.awi.de/en/expedition/ships/more-ships.html) at the research site 'Kabeltonne' (54° 11.3' N, 7° 54.0' E) from approximately 1 m water depth in 20 L carboys. The water samples for microbial biomass were subjected to a fractionating filtration directly upon arrival in the Biologische Anstalt Helgoland laboratory (typically less than one hour after sampling). The entire filtration process for all fractions was usually finished within 3 h, i.e. latest 4 h after the sampling. Further data collection was carried out by the staff of Biologische Anstalt Helgoland and MPI Bremen (chlorophyll A concentrations, total cell count), or Lennart Schada von Borzyskowski (glycolate concentrations and uptake rates, abundance of bhcC genes and transcripts), together with Peter Claus and Nina Socorro Cortina (glycolate concentrations).
Timing and spatial scale	Water samples were collected once daily in the morning (between 8 and 10 am) on each working day between March 1 and May 31, 2018. Water samples were collected in the North Sea close to Helgoland island, at the research site 'Kabeltonne' (54° 11.3' N, 7° 54.0' E), from approximately 1 m water depth.
Data exclusions	No data were excluded from the analyses.
Reproducibility	Not applicable, since sampling was carried out once daily.
Randomization	The water samples used in this study were selected and divided randomly in the different conditions to determine glycolate uptake rates. No criteria of selection were applied.
Blinding	Blinding of samples was not applicable for the kind of experiments included in this study, since water samples (in replicates) were taken on different dates and labeled accordingly.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	Water samples were collected once daily in the morning (between 8 and 10 am) on each working day between March 1 and May 31, independent of temperature or weather conditions.
Location	Water samples were collected in the North Sea close to Helgoland island, at the research site 'Kabeltonne' (54° 11.3' N, 7° 54.0' E), from approximately 1 m water depth.
Access and import/export	Water samples were collected without harming the habitat. No import or export of samples was conducted in the frame of this study.
Disturbance	No disturbance was caused by the study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Article

Bacteriophage targeting of gut bacterium attenuates alcoholic liver disease

https://doi.org/10.1038/s41586-019-1742-x

Received: 9 April 2019

Accepted: 2 October 2019

Published online: 13 November 2019

Yi Duan[1,2,29], Cristina Llorente[1,2,29], Sonja Lang[1], Katharina Brandl[3], Huikuan Chu[1], Lu Jiang[1,2], Richard C. White[4], Thomas H. Clarke[4], Kevin Nguyen[4], Manolito Torralba[5], Yan Shao[6], Jinyuan Liu[7], Adriana Hernandez-Morales[8], Lauren Lessor[9], Imran R. Rahman[10], Yukiko Miyamoto[1], Melissa Ly[11], Bei Gao[1], Weizhong Sun[1], Roman Kiesel[1], Felix Hutmacher[1], Suhan Lee[1], Meritxell Ventura-Cots[12], Francisco Bosques-Padilla[13], Elizabeth C. Verna[14], Juan G. Abraldes[15], Robert S. Brown Jr[16], Victor Vargas[17,18], Jose Altamirano[17], Juan Caballería[18,19], Debbie L. Shawcross[20], Samuel B. Ho[1,2], Alexandre Louvet[21], Michael R. Lucey[22], Philippe Mathurin[21], Guadalupe Garcia-Tsao[23,24], Ramon Bataller[12], Xin M. Tu[7], Lars Eckmann[1], Wilfred A. van der Donk[10,25,26], Ry Young[8,9], Trevor D. Lawley[6], Peter Stärkel[27], David Pride[1,11,28], Derrick E. Fouts[4] & Bernd Schnabl[1,2,28]*

Chronic liver disease due to alcohol-use disorder contributes markedly to the global burden of disease and mortality[1–3]. Alcoholic hepatitis is a severe and life-threatening form of alcohol-associated liver disease. The gut microbiota promotes ethanol-induced liver disease in mice[4], but little is known about the microbial factors that are responsible for this process. Here we identify cytolysin—a two-subunit exotoxin that is secreted by *Enterococcus faecalis*[5,6]—as a cause of hepatocyte death and liver injury. Compared with non-alcoholic individuals or patients with alcohol-use disorder, patients with alcoholic hepatitis have increased faecal numbers of *E. faecalis*. The presence of cytolysin-positive (cytolytic) *E. faecalis* correlated with the severity of liver disease and with mortality in patients with alcoholic hepatitis. Using humanized mice that were colonized with bacteria from the faeces of patients with alcoholic hepatitis, we investigated the therapeutic effects of bacteriophages that target cytolytic *E. faecalis*. We found that these bacteriophages decrease cytolysin in the liver and abolish ethanol-induced liver disease in humanized mice. Our findings link cytolytic *E. faecalis* with more severe clinical outcomes and increased mortality in patients with alcoholic hepatitis. We show that bacteriophages can specifically target cytolytic *E. faecalis*, which provides a method for precisely editing the intestinal microbiota. A clinical trial with a larger cohort is required to validate the relevance of our findings in humans, and to test whether this therapeutic approach is effective for patients with alcoholic hepatitis.

The most severe form of alcohol-related liver disease is alcoholic hepatitis; mortality ranges from 20% to 40% at 1–6 months, and as many as 75% of patients die within 90 days of a diagnosis of severe alcoholic hepatitis[7–9]. Therapy with corticosteroids is only marginally effective[9]. Early liver transplantation is the only curative therapy, but is offered only at select centres and to a limited group of patients[10].

Alcohol-related liver disease can be transmitted via faecal microbiota[4]. We investigated the microorganisms and microbial factors that are responsible for this transmissible phenotype and for progression of alcohol-related liver disease.

Cytolysin linked to increased mortality

We performed 16S ribosomal RNA (rRNA) gene sequencing to determine whether chronic alcohol use and alcoholic hepatitis are associated with an altered composition of the faecal microbiota. Differences in faecal microbiota composition were noted in patients with alcohol-use disorder and alcoholic hepatitis, compared to subjects without alcohol-use disorder (controls) (Fig. 1a, Extended Data Fig. 1a, b, Supplementary Tables 1, 2). One substantial difference that we observed was an increase in the proportion of *Enterococcus* spp. in patients with alcoholic hepatitis: in these patients, 5.59% of faecal bacteria were *Enterococcus* spp. compared with almost none in controls (0.023%; for comparison, 0.004% of all reads were *Enterococcus* spp. in the Human Microbiome Project) or patients with alcohol-use disorder (0.024%). Faecal samples from patients with alcoholic hepatitis had about 2,700-fold more *E. faecalis* than samples from controls, as measured by quantitative PCR (qPCR) (Extended Data Fig. 1c), which is consistent with the 16S rRNA sequencing results. About 80% of patients with alcoholic hepatitis are positive for *E. faecalis* in their faeces (Extended Data Fig. 1d).

A list of affiliations appears at the end of the paper.

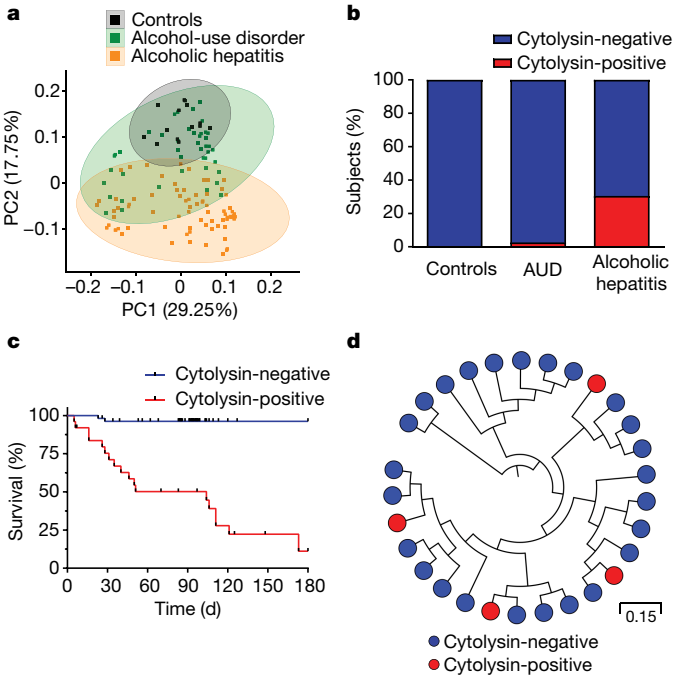


Fig. 1 | *E. faecalis* cytolysin is associated with mortality in patients with alcoholic hepatitis. a, 16S rRNA sequencing of faecal samples from controls ($n = 14$), patients with alcohol-use disorder ($n = 43$) or alcoholic hepatitis ($n = 75$). We use principal coordinate analysis (PCoA) based on Jaccard dissimilarity matrices to show β-diversity among groups at the genus level. The composition of faecal microbiota was significantly different between each group ($P < 0.01$). **b**, Percentage of subjects with faecal samples that were positive for both $cylL_L$ and $cylL_S$ DNA sequences (cytolysin-positive), in controls ($n = 25$), patients with alcohol-use disorder ($n = 38$) or alcoholic hepatitis ($n = 82$), assessed by qPCR. Statistically significant differences were detected between controls and patients with alcoholic hepatitis ($P < 0.01$), and between patients with alcohol-use disorder and patients with alcoholic hepatitis ($P < 0.001$). **c**, Kaplan–Meier curve of survival of patients with alcoholic hepatitis whose faecal samples were cytolysin-positive ($n = 25$) or cytolysin-negative ($n = 54$) ($P < 0.0001$). **d**, Core genome single-nucleotide polymorphism (SNP) tree of *E. faecalis* strains isolated from patients with alcoholic hepatitis ($n = 93$ strains, from 24 patients), showing phylogenetic diversity of cytolysin-positive (red) *E. faecalis*. Genomically identical isolates from the same patient were combined, and are shown as a single dot. Scale bar represents the nucleotide substitutions per SNP site. *P* values are determined by permutational multivariate analysis of variance (PERMANOVA) followed by false discovery rate (FDR) procedures (**a**), two-sided Fisher's exact test followed by FDR procedures (**b**) or two-sided log-rank (Mantel–Cox) test (**c**). The exact group size (n) and *P* values for each comparison are listed in Supplementary Table 10.

The colonization of mice with *E. faecalis* induces mild hepatic steatosis and exacerbates ethanol-induced liver disease[11], by mechanisms that are unclear. Cytolysin is a bacterial exotoxin (or bacteriocin) that is produced by *E. faecalis*[12], and which contains two post-translationally modified peptides ($CylL_L''$ and $CylL_S''$) in its bioactive form[6]. The two peptides are encoded by two separate genes: $cylL_L$ and $cylL_S$, respectively[12]. Cytolysin has lytic activity against not only Gram-positive bacteria, but also eukaryotic cells[13]. We detected $cylL_L$ and $cylL_S$ genomic DNA (cytolysin-positive) in faecal samples from 30% of patients with alcoholic hepatitis; none of the faecal samples from controls and only one sample from a patient with alcohol-use disorder was cytolysin-positive, as detected by qPCR (Fig. 1b). Importantly, 89% of cytolysin-positive patients with alcoholic hepatitis died within 180 days of admission, compared to only 3.8% of cytolysin-negative patients ($P < 0.0001$) (Fig. 1c). Among the cytolysin-positive patients, 72.2% (13 out of 18) died owing to liver failure (including complications related to liver failure, such as gastrointestinal bleeding) (Supplementary Table 2). Infection was not associated with 30-day, 90-day or 180-day mortality ($P = 0.403$, 0.234 or 0.098) in patients with alcoholic hepatitis.

Our univariate logistic and Cox regression of laboratory and clinical parameters found an association between the detection of cytolysin-encoding genes in faeces and the international normalized ratio (INR), platelet count, the model for end-stage liver disease (MELD) score, the sodium MELD score, the age, serum bilirubin, INR and serum creatinine (ABIC) score and death (Supplementary Table 3). In the multivariate Cox analysis, detection of cytolysin-encoding genes in faeces was associated with 90-day ($P = 0.004$) and with 180-day mortality ($P = 0.001$) (Supplementary Table 3), even after we adjusted for the geographical origin of the patient, antibiotic treatment, platelet count, and creatinine, bilirubin and INR as components of the MELD score. We found no multicollinearity between the detection of faecal cytolysin-encoding genes and these cofactors (variance inflation factor < 1.6), which indicates that cytolysin is an independent predictor of mortality in patients with alcoholic hepatitis. When we performed receiver-operating-characteristic curve analysis for 90-day mortality, cytolysin had an area under the curve of 0.81, which was superior to other widely used predictors for mortality in clinical practice (Extended Data Fig. 1e). On the basis of our findings, we propose that the detection of cytolysin may be a prognostic factor for more severe liver-related outcomes and increased risk of death, and a stronger predictor of mortality than MELD, ABIC and the discriminant function score.

To determine phylogeny of *E. faecalis* in patients with alcoholic hepatitis, we performed targeted culturing from stool samples. Whole-genome sequencing of 93 *E. faecalis* isolates revealed a broad phylogenetic diversity of cytolysin-positive *E. faecalis* from patients with alcoholic hepatitis (Fig. 1d), which indicates that cytolysin production is a variable trait among *E. faecalis* isolates and that cytolysin is carried in mobile genetic elements, which include both chromosomally encoded pathogenicity islands and plasmids[14]. Detection of any other antimicrobial resistance genes or virulence genes in *E. faecalis* isolates did not correlate with disease severity or mortality in patients with alcoholic hepatitis (Supplementary Table 4).

The total amount of faecal *E. faecalis*, or faecal *E. faecalis* positivity, did not correlate with disease severity or mortality in patients with alcoholic hepatitis (Supplementary Tables 5, 6). Cytolysin-positive and cytolysin-negative patients with alcoholic hepatitis had similar amounts of faecal *E. faecalis* (Extended Data Fig. 1f). Although there were differences in the composition of the gut microbiota in patients with alcoholic hepatitis from different geographical regions (Extended Data Fig. 1g), the proportion of cytolysin-positive patients, total amount of faecal *E. faecalis*, faecal *E. faecalis* positivity (Extended Data Fig. 1h–j), treatment and clinical outcomes (30-day and 90-day mortality) did not differ significantly among the regions or centres (Supplementary Table 7). In addition, cirrhosis was not associated with cytolysin positivity, the total amount of faecal *E. faecalis* or faecal *E. faecalis* positivity in patients with alcoholic hepatitis (Extended Data Fig. 1k–m, Supplementary Tables 4–6). These results confirm our findings that the presence of cytolysin-producing *E. faecalis* rather than the total amount or presence of *E. faecalis* per se determines the severity of alcoholic hepatitis and mortality.

Cytolysin and ethanol-induced liver disease

To determine whether cytolysin contributes to liver damage mediated by *E. faecalis*, we gavaged mice with a cytolytic *E. faecalis* strain (FA2-2(pAM714)) or a non-cytolytic *E. faecalis* strain (FA2-2(pAM771))[5]; the mice were then placed on a chronic–binge ethanol diet[15]. Compared to mice gavaged with phosphate-buffered saline (PBS), mice fed with ethanol after they were gavaged with cytolytic *E. faecalis* developed more severe liver injury as indicated by a higher level of alanine aminotransferase (ALT) (Extended Data Fig. 2a) and increased hepatic steatosis (Extended Data Fig. 2b, c). Mice that were fed ethanol after they

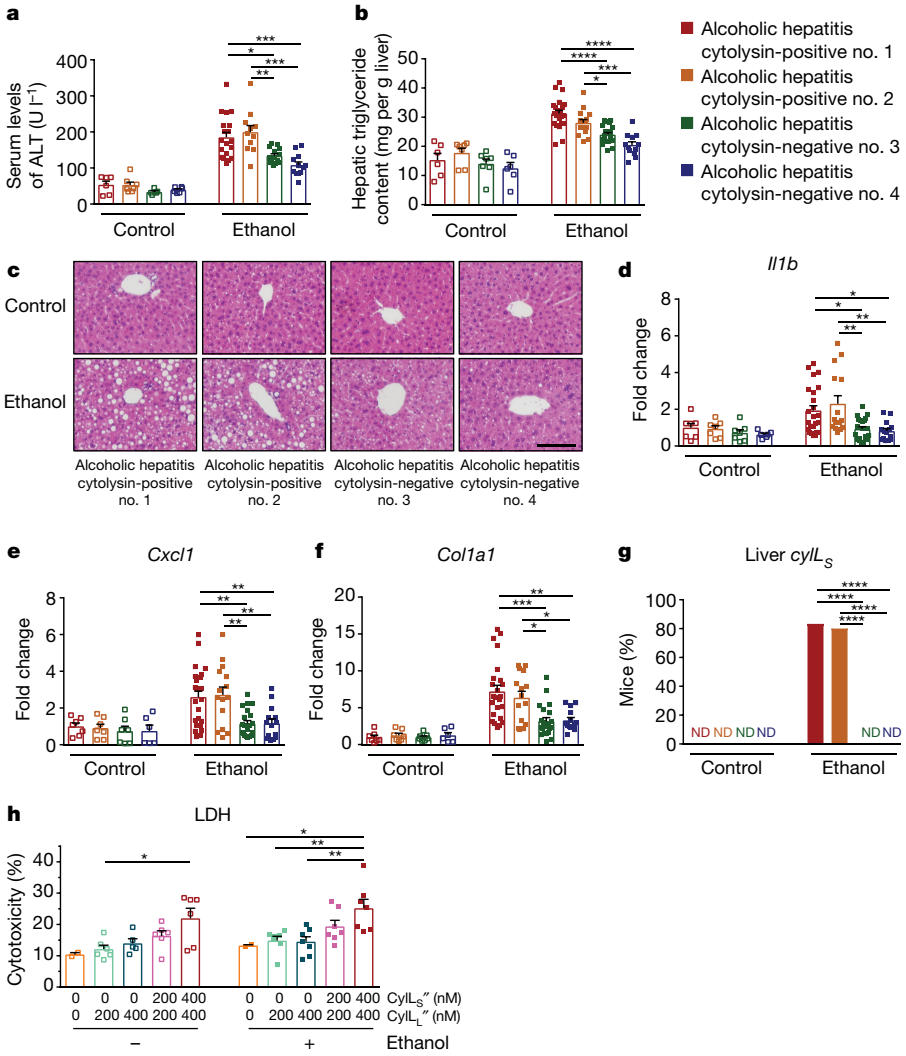


Fig. 2 | Transplantation of faeces from cytolysin-positive patients with alcoholic hepatitis exacerbates ethanol-induced liver disease in gnotobiotic mice. a–**g**, C57BL/6 germ-free mice were colonized with faeces from two cytolysin-positive and two cytolysin-negative patients with alcoholic hepatitis, and subjected to the chronic–binge feeding model. **a**, Serum levels of ALT. **b**, Hepatic triglyceride content. **c**, Representative sections of liver stained with haematoxylin and eosin (H & E). **d**–**f**, Hepatic levels of mRNAs that encode *Il1b*, *Cxcl1* and *Col1a1*. **g**, Proportions of mice that were positive for cytolysin in the liver, measured by qPCR for *cylL$_S$*. **h**, Lactate dehydrogenase (LDH) assay to measure cytotoxicity of hepatocytes isolated from mice that were fed an oral isocaloric control diet (five groups, left) or chronic–binge ethanol diet (five groups, right), and incubated with vehicle, CylL$_S$″, CylL$_L$″ or both of the cytolysin subunits at the indicated concentrations without (–) or with (+) ethanol (25 mM) for 3 h. The survival of hepatocytes was determined in three independent experiments. Scale bar, 100 μm. Results are expressed as mean ± s.e.m. (**a**, **b**, **d**–**f**, **h**). *P* values are determined by one-way analysis of variance (ANOVA) with Tukey's post hoc test (**a**, **b**, **d**–**f**), two-sided Fisher's exact test followed by FDR procedures (**g**) or two-way ANOVA with Tukey's post hoc test (**h**). All results were generated from at least three independent replicates. The exact group size (*n*) and *P* values for each comparison are listed in Supplementary Table 10. **P* < 0.05, ***P* < 0.01, ****P* < 0.001, *****P* < 0.0001.

were gavaged with cytolytic *E. faecalis* also had more liver inflammation with higher expression levels of mRNAs that encode inflammatory cytokines and chemokines (*Il1b*, *Cxcl1* and *Cxcl2*) (Extended Data Fig. 2d–f), compared with mice given PBS. Mice that were fed ethanol after they were gavaged with non-cytolytic *E. faecalis* had significantly less ethanol-induced liver injury, steatosis and inflammation (Extended Data Fig. 2a–f) and longer survival times (Extended Data Fig. 2g), as compared with mice that were fed ethanol after they were administered with cytolytic *E. faecalis*.

To explore the mechanism of cytolysin-associated liver damage, we measured cytolysin in the liver. *CylL$_S$* was significantly increased in the liver of mice given cytolytic *E. faecalis* but not in the liver of mice that were not given *E. faecalis* or of mice gavaged with non-cytolytic *E. faecalis* after chronic ethanol administration (Extended Data Fig. 2h). *E. faecalis* was detectable in the liver of mice given cytolytic and non-cytolytic *E. faecalis* and fed an ethanol diet, but not when mice were fed an isocaloric (control) diet (Extended Data Fig. 2i); this indicates that ethanol-induced changes in the gut barrier are necessary for the translocation of cytolytic *E. faecalis* from the intestine to the liver. The livers of ethanol-fed mice that were given cytolytic or non-cytolytic *E. faecalis* had positive *E. faecalis* cultures (Extended Data Fig. 2j). We observed an increased intestinal permeability in ethanol-fed mice compared with mice fed with an isocaloric diet, but this was independent of gavaging cytolytic or non-cytolytic *E. faecalis* after chronic ethanol administration (Extended Data Fig. 2k), indicating that cytolysin does not affect intestinal barrier function.

Administration of cytolytic or non-cytolytic *E. faecalis* to mice did not significantly change the composition of the intestinal microbiota, as shown by 16S rRNA gene sequencing (Extended Data Fig. 2l). Cytolytic *E. faecalis* did not affect intestinal absorption or hepatic metabolism

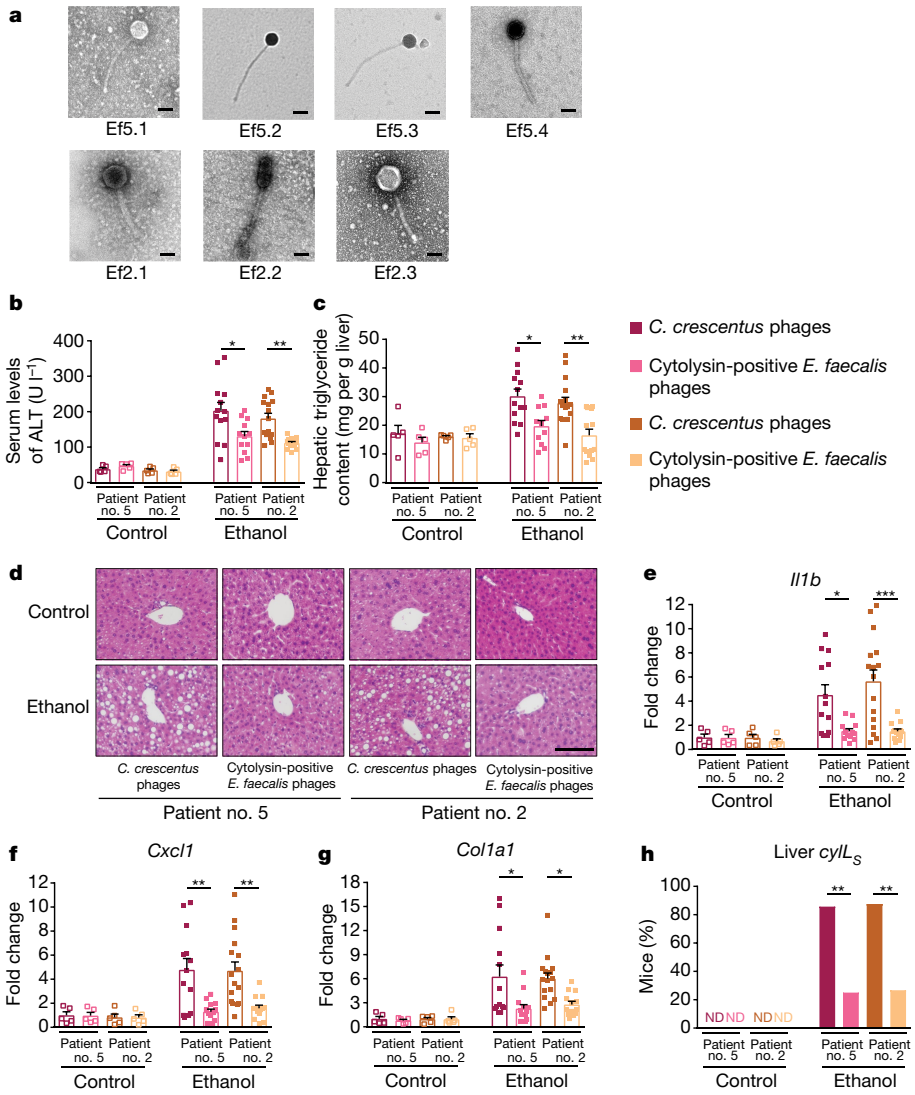


Fig. 3 | Phage therapy against cytolytic *E. faecalis* abolishes ethanol-induced liver disease in gnotobiotic mice. a, Transmission electron microscopy revealed that the phages we isolated were either siphophages (Ef5.1, Ef5.2, Ef5.3, Ef5.4 and Ef2.2) or myophages (Ef2.1 and Ef2.3). Scale bar, 50 nm. **b**–**h**, C57BL/6 germ-free mice were colonized with faeces from two cytolysin-positive patients with alcoholic hepatitis (faeces from one of these patients were also used in Fig. 2) and subjected to the chronic–binge feeding model, gavaged with control phages against *C. crescentus* (10^{10} plaque-forming units (PFUs)) or a cocktail of three or four different phages that target cytolytic *E. faecalis* (10^{10} PFUs), 1 day before an ethanol binge. **b**, Serum levels of ALT. **c**, Hepatic triglyceride content. **d**, Representative H & E-stained liver sections. Scale bar, 100 µm. **e**–**g**, Hepatic levels of mRNAs that encode *Il1b*, *Cxcl1* and *Col1a1*. **h**, Proportions of mice that were positive for cytolysin in the liver, measured by qPCR for $cylL_S$. Results are expressed as mean ± s.e.m. (**b**, **c**, **e**–**g**). *P* values are determined by two-way ANOVA with Tukey's post hoc test (**b**, **c**, **e**–**g**) or two-sided Fisher's exact test followed by FDR procedures (**h**). All results are generated from at least three independent replicates. The exact group size (*n*) and *P* values for each comparison are listed in Supplementary Table 10. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

of ethanol, as indicated by serum levels of ethanol and hepatic levels of *Adh1* and *Cyp2e1* mRNAs (which encode the two primary enzymes that metabolize ethanol in the liver) (Extended Data Fig. 2m, n). These results indicate that *E. faecalis* that produce cytolysin promote ethanol-induced liver disease in mice.

To extend our findings to humans, we colonized germ-free mice with faeces from cytolysin-positive and cytolysin-negative patients with alcoholic hepatitis (Supplementary Table 8). Consistent with our findings from mice colonized with cytolytic *E. faecalis*, gnotobiotic C57BL/6 mice colonized with faeces from two cytolysin-positive patients developed more severe ethanol-induced liver injury, steatosis, inflammation and fibrosis than mice given faeces from two cytolysin-negative patients (Fig. 2a–f, Extended Data Fig. 3a–d). Transplantation of faeces from cytolysin-positive patients reduced the survival time of the mice (Extended Data Fig. 3e) and increased translocation of cytolytic *E. faecalis* to the liver after ethanol administration (Fig. 2g). The overall composition of the intestinal microbiota was not different between mice fed the control diet and colonized with faeces from cytolysin-positive or cytolysin-negative donors with alcoholic hepatitis, as shown by 16S rRNA gene sequencing. Mice transplanted with faeces from one of the cytolysin-positive patients with alcoholic hepatitis (patient no. 2) showed a microbiota that was significantly different from that of the other mouse groups after ethanol administration (Extended Data Fig. 3f). Non-cytolytic *E. faecalis* was not detected in stool samples from donors with cytolytic *E. faecalis* (Extended Data Fig. 3g). We did not observe differences in intestinal absorption or hepatic metabolism of ethanol between mice colonized with faeces from cytolysin-positive versus cytolysin-negative patients (Extended Data Fig. 3h, i). Together, these results provide further evidence that cytolysin promotes ethanol-induced liver disease.

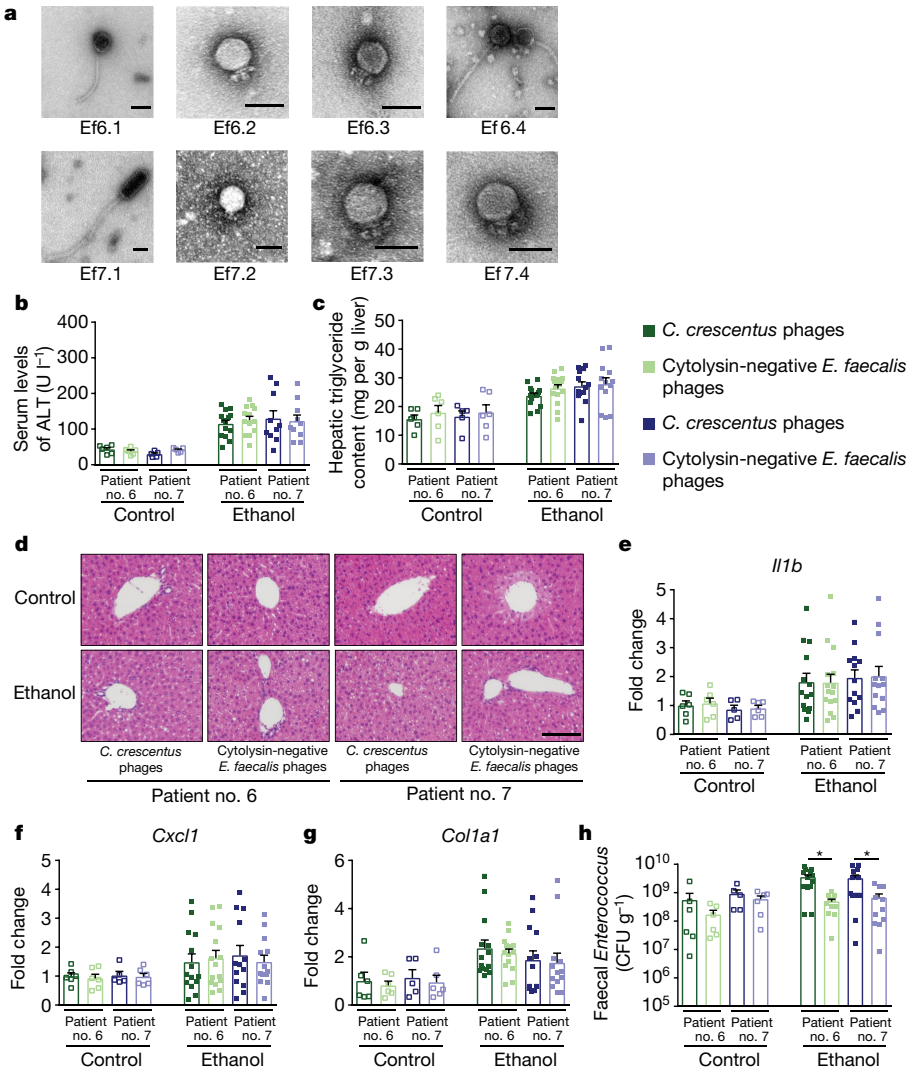


Fig. 4 | Phages that target non-cytolytic *E. faecalis* do not reduce ethanol-induced liver disease in gnotobiotic mice. a, Transmission electron microscopy revealed that the phages we isolated were either podophages (Ef6.2, Ef6.3, Ef7.2, Ef7.3 and Ef7.4) or siphophages (Ef6.1, Ef6.4 and Ef7.1). Scale bar, 50 nm. **b**–**h**, C57BL/6 germ-free mice were colonized with faeces from two cytolysin-negative patients with alcoholic hepatitis and subjected to the chronic–binge feeding model, gavaged with control phages against *C. crescentus* (10^{10} PFUs) or a cocktail of four different phages that target non-cytolytic *E. faecalis* (10^{10} PFUs), 1 day before an ethanol binge. **b**, Serum levels of ALT. **c**, Hepatic triglyceride content. **d**, Representative H & E-stained liver sections. Scale bar, 100 μm. **e**–**g**, Hepatic levels of mRNAs that encode *Il1b*, *Cxcl1* and *Col1a1*. **h**, Faecal colony-forming units (CFUs) of *Enterococcus*. Results are expressed as mean ± s.e.m. (**b**, **c**, **e**–**h**). *P* values are determined by two-way ANOVA with Tukey's post hoc test (**b**, **c**, **e**–**h**). All results were generated from at least three independent replicates. The exact group size (*n*) and *P* values for each comparison are listed in Supplementary Table 10. $^{*}P < 0.05$.

To determine the mechanism by which cytolysin increases liver disease, we isolated hepatocytes from mice fed ethanol or control diets, and stimulated them with pure bioactive cytolysin peptides ($CylL_L''$ and $CylL_S'')$[6]. Incubation of the primary mouse hepatocytes with the two cytolysin subunits caused a dose-dependent increase in cell death compared to hepatocytes that were incubated with vehicle or with one subunit only (Fig. 2h). When we isolated hepatocytes from ethanol-fed mice and then incubated these hepatocytes with ethanol, we did not observe increased levels of cytolysin-induced cell death compared to hepatocytes isolated from mice on the control diet, which indicates that cytolysin-induced hepatocyte cell death was independent of ethanol. The cytotoxic effects of cytolysin are possibly mediated by pore formation, resulting in cell lysis[14].

Bacteriophage treatment in liver disease

To further demonstrate the potential causative role of cytolytic *E. faecalis* for the development of ethanol-induced steatohepatitis, we investigated the effects of treatment with bacteriophages (hereafter, phages). Phages are ubiquitous in bacteria-rich environments, including the gut[16]. *E. faecalis* phages that are highly strain-specific can be isolated[17], which potentially makes the direct editing of gut microbiota feasible. It has previously been shown that *Atp4a*$^{SI/SI}$ mice, which lack gastric acid, have overgrowth of intestinal enterococci, which is associated with increased susceptibility to alcohol-induced steatohepatitis[11]. The gavaging of wild-type mice with an *E. faecalis* strain isolated from *Atp4a*$^{SI/SI}$ mice led to increased ethanol-induced steatohepatitis[11]. We found that this same *E. faecalis* strain expressed cytolysin. We then isolated four distinct phages from sewage water. These phages lyse the cytolytic *E. faecalis* strain isolated from *Atp4a*$^{SI/SI}$ mice. All four phages were podophages of the virulent *Picovirinae* group (Extended Data Fig. 4). *Atp4a*$^{SI/SI}$ mice and their wild-type littermates were then placed on the chronic–binge ethanol diet and gavaged with the lytic phage cocktail. Phages directed against *Caulobacter crescentus*, a bacterium that is present in freshwater lakes and streams[18] but that does not colonize humans or rodents[19,20], were used as controls.

Compared to *Atp4a*$^{Sl/Sl}$ mice gavaged with control phages or vehicle, *Atp4a*$^{Sl/Sl}$ mice gavaged with phages that target cytolytic *E. faecalis* had less severe liver injury, steatosis and inflammation after chronic ethanol feeding (Extended Data Fig. 5a–f). Administration of *E. faecalis* phages significantly reduced levels of cytolysin in the liver (Extended Data Fig. 5g) as well as faecal amounts of *Enterococcus* (Extended Data Fig. 5h). Phage administration did not affect the overall composition of the faecal microbiome, intestinal absorption or hepatic metabolism of ethanol (Extended Data Fig. 5i–k).

To develop a therapeutic approach to precisely edit the intestinal microbiota, we cultured cytolytic *E. faecalis* strains from the faecal samples of patients with alcoholic hepatitis. We then isolated lytic phages from sewage water against these cytolytic *E. faecalis* strains; these phages had siphophage or myophage morphology (Fig. 3a, Extended Data Fig. 6). Gnotobiotic mice were colonized with faeces from two cytolysin-positive patients with alcoholic hepatitis (Supplementary Table 8) and given three or four different—but patient-specific—lytic phages against cytolytic *E. faecalis*. The phages against cytolytic *E. faecalis* abolished ethanol-induced liver injury and steatosis, as shown by lower levels of ALT, lower percentages of hepatic cells positive for terminal deoxynucleotide transferase-mediated dUTP nick-end labelling, and lower levels of hepatic triglycerides and oil red O-staining (Fig. 3b–d, Extended Data Fig. 7a, b), as well as by decreased hepatic levels of *Il1b*, *Cxcl1*, *Cxcl2*, *Col1a1* and *Acta2* mRNAs, and reduced hepatic levels of *cyllL*$_S$, as compared with mice given control phages (against *C. crescentus*) (Fig. 3e–h, Extended Data Fig. 7c, d). Treatment with phages against cytolytic *E. faecalis* also reduced faecal amounts of *Enterococcus* (Extended Data Fig. 7e) without affecting the overall composition of the gut microbiota (Extended Data Fig. 7f). Intestinal absorption of ethanol and hepatic metabolism were similar in all groups (Extended Data Fig. 7g, h).

To demonstrate that the effect of phage treatment occurs via the targeting of cytolysin-positive *E. faecalis*, rather than a reduction in cytolysin-negative *E. faecalis*, we colonized gnotobiotic mice with faeces from cytolysin-negative patients with alcoholic hepatitis (Supplementary Table 8). Phages against non-cytolytic *E. faecalis* from patients were isolated from sewage water; they had siphophage or podophage morphology (Fig. 4a, Extended Data Fig. 8). These phages did not reduce features of ethanol-induced liver disease compared with control phages (Fig. 4b–g, Extended Data Fig. 9a–h), despite the reduction of faecal *Enterococcus* (Fig. 4h). Our findings indicate that treatment with lytic phages can selectively attenuate the ethanol-induced liver disease caused by cytolysin-positive *E. faecalis* in humanized mice.

Discussion

Phage-based therapies have predominantly been studied in patients with bacterial infections in the gastrointestinal tract[21–23], urinary tract[24,25] and other organ systems[26–28]. The results of these studies—although mixed in terms of efficacy—strongly suggest that phage treatment offers a safe alternative to antibiotics[26,27]. However, safety studies are required for complex populations (such as patients with alcoholic hepatitis), because phages can induce a strong immune reaction[29]. Further work is required to determine whether phages that target cytolytic *E. faecalis* might be used to treat patients with alcoholic hepatitis, a life-threatening disease that at present has no effective treatment. Eradication of this specific bacterial strain might produce better outcomes than current treatments, and environmental sources can be used to easily isolate phages that target cytolysin-positive *E. faecalis*. Here we provide an example of the efficacy of approaches based on phages in mice for a disease that is not considered a classic infectious disease. Our data also suggest that cytolysin may be used as a predictive biomarker of severe alcoholic hepatitis; an independent, prospective cohort is therefore needed to validate cytolysin as a biomarker, and to extend the phage findings in mice to human patients.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1742-x.

1. Lozano, R. et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**, 2095–2128 (2012).
2. Lee, B. P., Vittinghoff, E., Dodge, J. L., Cullaro, G. & Terrault, N. A. National trends and long-term outcomes of liver transplant for alcohol-associated liver disease in the United States. *JAMA Intern. Med.* **179**, 340–348 (2019).
3. Rehm, J. et al. Burden of disease associated with alcohol use disorders in the United States. *Alcohol. Clin. Exp. Res.* **38**, 1068–1077 (2014).
4. Llopis, M. et al. Intestinal microbiota contributes to individual susceptibility to alcoholic liver disease. *Gut* **65**, 830–839 (2016).
5. Ike, Y., Clewell, D. B., Segarra, R. A. & Gilmore, M. S. Genetic analysis of the pAD1 hemolysin/bacteriocin determinant in *Enterococcus faecalis*: Tn*917* insertional mutagenesis and cloning. *J. Bacteriol.* **172**, 155–163 (1990).
6. Tang, W. & van der Donk, W. A. The sequence of the enterococcal cytolysin imparts unusual lanthionine stereochemistry. *Nat. Chem. Biol.* **9**, 157–159 (2013).
7. Maddrey, W. C. et al. Corticosteroid therapy of alcoholic hepatitis. *Gastroenterology* **75**, 193–199 (1978).
8. Dominguez, M. et al. A new scoring system for prognostic stratification of patients with alcoholic hepatitis. *Am. J. Gastroenterol.* **103**, 2747–2756 (2008).
9. Thursz, M. R. et al. Prednisolone or pentoxifylline for alcoholic hepatitis. *N. Engl. J. Med.* **372**, 1619–1628 (2015).
10. Mathurin, P. & Lucey, M. R. Management of alcoholic hepatitis. *J. Hepatol.* **56**, S39–S45 (2012).
11. Llorente, C. et al. Gastric acid suppression promotes alcoholic liver disease by inducing overgrowth of intestinal *Enterococcus*. *Nat. Commun.* **8**, 837 (2017).
12. Gilmore, M. S. et al. Genetic structure of the *Enterococcus faecalis* plasmid pAD1-encoded cytolytic toxin system and its relationship to lantibiotic determinants. *J. Bacteriol.* **176**, 7335–7344 (1994).
13. Cox, C. R., Coburn, P. S. & Gilmore, M. S. Enterococcal cytolysin: a novel two component peptide system that serves as a bacterial defense against eukaryotic and prokaryotic cells. *Curr. Protein Pept. Sci.* **6**, 77–84 (2005).
14. Van Tyne, D., Martin, M. J. & Gilmore, M. S. Structure, function, and biology of the *Enterococcus faecalis* cytolysin. *Toxins* **5**, 895–911 (2013).
15. Bertola, A., Mathews, S., Ki, S. H., Wang, H. & Gao, B. Mouse model of chronic and binge ethanol feeding (the NIAAA model). *Nat. Protoc.* **8**, 627–637 (2013).
16. Ogilvie, L. A. & Jones, B. V. The human gut virome: a multifaceted majority. *Front. Microbiol.* **6**, 918 (2015).
17. Chatterjee, A. et al. Bacteriophage resistance alters antibiotic-mediated intestinal expansion of enterococci. *Infect. Immun.* **87**, e00085-19 (2019).
18. Poindexter, J. S. Biological properties and classification of the *Caulobacter* group. *Bacteriol. Rev.* **28**, 231–295 (1964).
19. Shin, J. et al. Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. *Sci. Rep.* **6**, 29681 (2016).
20. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
21. Marcuk, L. M. et al. Clinical studies of the use of bacteriophage in the treatment of cholera. *Bull. World Health Organ.* **45**, 77–83 (1971).
22. Sarker, S. A. et al. Oral phage therapy of acute bacterial diarrhea with two coliphage preparations: a randomized trial in children from Bangladesh. *EBioMedicine* **4**, 124–137 (2016).
23. Dalmasso, M., Hill, C. & Ross, R. P. Exploiting gut bacteriophages for human health. *Trends Microbiol.* **22**, 399–405 (2014).
24. Ujmajuridze, A. et al. Adapted bacteriophages for treating urinary tract infections. *Front. Microbiol.* **9**, 1832 (2018).
25. Khawaldeh, A. et al. Bacteriophage therapy for refractory *Pseudomonas aeruginosa* urinary tract infection. *J. Med. Microbiol.* **60**, 1697–1700 (2011).
26. Schooley, R. T. et al. Development and use of personalized bacteriophage-based therapeutic cocktails to treat a patient with a disseminated resistant *Acinetobacter baumannii* infection. *Antimicrob. Agents Chemother.* **61**, e00954-17 (2017).
27. Dedrick, R. M. et al. Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant *Mycobacterium abscessus*. *Nat. Med.* **25**, 730–733 (2019).
28. Fish, R. et al. Compassionate use of bacteriophage therapy for foot ulcer treatment as an effective step for moving toward clinical trials. *Methods Mol. Biol.* **1693**, 159–170 (2018).
29. Górski, A. et al. Phages and immunomodulation. *Future Microbiol.* **12**, 905–914 (2017).

[1]Department of Medicine, University of California San Diego, La Jolla, CA, USA. [2]Department of Medicine, VA San Diego Healthcare System, San Diego, CA, USA. [3]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA. [4]J. Craig Venter Institute, Rockville, MD, USA. [5]J. Craig Venter Institute, La Jolla, CA, USA. [6]Host–Microbiota Interactions Laboratory, Wellcome Sanger Institute, Hinxton, UK. [7]Division of Biostatistics and Bioinformatics, Department of Family Medicine and Public Health, University of California San Diego, La Jolla, CA, USA. [8]Department of Biochemistry and Biophysics, Texas A & M University, College Station, TX, USA. [9]Center for Phage Technology, Texas A & M AgriLife Research and Texas A & M University, College Station, TX, USA. [10]Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [11]Department of Pathology, University of California San Diego, La Jolla, CA, USA. [12]Division of Gastroenterology, Hepatology and Nutrition, Department of Medicine, University of Pittsburgh Medical Center, Pittsburgh Liver Research Center, Pittsburgh, PA, USA. [13]Hospital Universitario, Departamento de Gastroenterología, Universidad Autonoma de Nuevo Leon, Monterrey, Mexico. [14]Division of Digestive and Liver Diseases, Department of Medicine, Columbia University College of Physicians and Surgeons, New York, NY, USA. [15]Department of Medicine, University of Alberta, Edmonton, Alberta, Canada. [16]Division of Gastroenterology and Hepatology, Weill Cornell Medical College, New York, NY, USA. [17]Liver Unit, Hospital Universitari Vall d'Hebron, Universitat Autònoma de Barcelona, Barcelona, Spain. [18]Centro de Investigación en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Barcelona, Spain. [19]Liver Unit, Hospital Clinic, Barcelona, Spain. [20]Liver Sciences, Department of Inflammation Biology, School of Infectious Diseases and Microbial Sciences, King's College London, London, UK. [21]Service des Maladies de L'appareil Digestif et Unité INSERM, Hôpital Huriez, Lille, France. [22]Division of Gastroenterology and Hepatology, Department of Medicine, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA. [23]Section of Digestive Diseases, Yale University School of Medicine, New Haven, CT, USA. [24]Section of Digestive Diseases, VA Connecticut Healthcare System, West Haven, CT, USA. [25]Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [26]Howard Hughes Medical Institute, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [27]St Luc University Hospital, Université Catholique de Louvain, Brussels, Belgium. [28]Center for Innovative Phage Applications and Therapeutics, University of California San Diego, La Jolla, CA, USA. [29]These authors contributed equally: Yi Duan, Cristina Llorente. *e-mail: beschnabl@ucsd.edu

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Patient cohorts

Patient cohorts have previously been described[30–32]. We evaluated 26 subjects without alcohol-use disorder (controls; social drinkers consuming less than 20 g/day), 44 patients with alcohol-use disorder and 88 patients with alcoholic hepatitis. Patients with alcohol-use disorder fulfilling the DSM IV criteria[33] of alcohol dependence and with active alcohol consumption (self-reported >60 g/day) presented with various stages of liver disease (21% had advanced F3/4 fibrosis based on fibrosis-4 index) (Supplementary Table 1). Patients with alcohol-use disorder were recruited from an alcohol withdrawal unit in San Diego and Brussels, where they followed a detoxification and rehabilitation programme. At admission to the hospital, a complete medication and medical history was taken, and a complete physical examination was performed, including collection of bio-specimens, basic demographic data (such as age, gender, weight and height) and self-reported daily alcohol consumption. Patients were actively drinking until the day of admission. Controls or patients with alcohol-use disorder did not take antibiotics or immunosuppressive medication during the two months preceding enrolment. Other exclusion criteria were diabetes, inflammatory bowel disease, known liver disease of any other aetiology, and clinically important cardiovascular, pulmonary or renal co-morbidities. Patients with alcoholic hepatitis were enrolled from the InTeam Consortium (ClinicalTrials.gov identifier number: NCT02075918) from centres in the USA, Mexico, UK, France and Spain. Inclusion criteria for this study were active alcohol abuse (>50 g/day for men and >40 g/day for women) in the past 3 months, aspartate aminotransferase (AST) > ALT and total bilirubin >3 mg/dl in the past 3 months, and a liver biopsy and/or clinical picture consistent with alcoholic hepatitis. Exclusion criteria were autoimmune liver disease (ANA >1/320), chronic viral hepatitis, hepatocellular carcinoma, complete portal vein thrombosis, extrahepatic terminal disease, pregnancy and a lack of signed informed consent. In all patients, the clinical picture was consistent with alcoholic hepatitis and in patients who underwent liver biopsy, the histology was consistent with the diagnosis of alcoholic hepatitis. Liver biopsies were only done if clinically indicated as part of routine clinical care for diagnostic purposes of alcoholic hepatitis. Bio-specimens were collected during their admission to the hospital. The median time of specimen collection was 4 days following admission to the hospital (range 0–24, $n = 82$). For one patient who underwent liver transplantation, the transplantation date was considered as date of death. Patients were censored at the time point at which they were last seen alive. The baseline characteristics are shown in Supplementary Tables 1, 2. Faecal 16S rRNA sequencing, *Enterococcus* culture and qPCR were performed. The MELD score, ABIC score and discriminant function were calculated from all alcoholic hepatitis patients from whom respective laboratory values were available. The protocol was approved by the Ethics Committee of Hôpital Huriez, Universidad Autonoma de Nuevo Leon, Hospital Universitari Vall d'Hebron, King's College London, Yale University, University of North Carolina at Chapel Hill, Weill Cornell Medical College, Columbia University, University of Wisconsin, VA San Diego Healthcare System, University of California San Diego (UCSD) and Université Catholique de Louvain. Patients were enrolled after written informed consent was obtained from each patient.

Mice

C57BL/6 mice were purchased from Charles River and used in Fig. 2h and Extended Data Fig. 2. C57BL/6 germ-free mice were bred at UCSD and used in Figs. 2a–g, 3, 4, Extended Data Figs. 3, 7 and 9. Sublytic *Atp4a*$^{Sl/Sl}$ mice on a C57BL/6 background have previously been described[11,34] and heterozygous mice were used for breeding; sublytic *Atp4a*$^{Sl/Sl}$ littermate mice and their wild-type littermates were used in Extended Data Fig. 5.

Female and male mice (age of 9–12 weeks) were placed on a chronic–binge ethanol diet (NIAAA model) as previously described[15]. Mice were fed with Lieber–DeCarli diet and the caloric intake from ethanol was 0% on days 1–5 and 36% from day 6 until the end of the study period. At day 16, mice were gavaged with a single dose of ethanol (5 g/kg body weight) in the early morning and killed 9 h later. Pair-fed control mice received a diet with an isocaloric substitution of dextrose.

Stool samples from patients with alcoholic hepatitis (Fig. 1) were used for faecal transplantation in germ-free mice. Mice were gavaged with 100 µl of stool samples (1 g stool dissolved in 30 ml Luria–Bertani (LB) medium containing 15% glycerol under anaerobic conditions), starting at an age of 5–6 weeks and repeated 2 weeks later. Two weeks after the second gavage, mice were placed on the ethanol or control (isocaloric) diet.

In studies of the effects of cytolysin, 5×10^8 CFUs of a cytolytic *E. faecalis* strain (FA2-2(pAM714)), a non-cytolytic *E. faecalis* strain (FA2-2(pAM771))[5] (*E. faecalis* Δcytolysin) (kindly provided by M. S. Gilmore), or PBS (vehicle control) were fed to mice by gavage every third day, starting from day 6 through day 15 of ethanol feeding. Administration every third day was necessary, given that *E. faecalis* does not colonize mice[11] (Extended Data Fig. 2o). To determine the effect of phage treatment, 10^{10} PFUs of *E. faecalis* phages (or *C. crescentus* phage phiCbK as control)[35] were gavaged to the mice 1 day before the ethanol binge (at day 16). All animal studies were reviewed and approved by the Institutional Animal Care and Use Committee of UCSD.

Bacteriophage isolation and amplification

The *E. faecalis* strain from *Atp4a*$^{Sl/Sl}$ mice faeces has previously been isolated[11] and was used to isolate phages Efmus1, Efmus2, Efmus3 and Efmus4 (phages specific to the *E. faecalis* strain isolated from mouse faeces were named as Efmus with a number (Ef for *E. faecalis*, mus for mouse, digit for isolation order). *E. faecalis* strains from human stool samples were isolated using methods described below, and the corresponding phages were named as Ef with patient number plus a digit (Ef for *E. faecalis*, last digit for isolation order). All *E. faecalis* strains were grown statically in brain–heart infusion (BHI) broth or on BHI agar at 37 °C. *C. crescentus* phage phiCbK was purified as previously described[35].

E. faecalis phages were isolated from untreated raw sewage water obtained from North City Water Reclamation Plant in San Diego. Fifty millilitres of raw sewage water was centrifuged at 8,000*g* for 1 min at room temperature to pellet large particles. The supernatant was passed through a 0.45-µm and then a 0.2-µm syringe filter (Whatman, PES membrane). One hundred microlitres of the clarified sewage was mixed with 100 µl overnight *E. faecalis* culture and then added to BHI broth top agar (0.5% agar) and poured over a BHI plate (1.5% agar). After overnight growth at 37 °C, the resulting plaques were recovered using a sterile pipette tip in 500 µl PBS. Phages were replaqued on *E. faecalis* three more times to ensure that the phages were clonal isolates.

High-titre phage stocks were propagated by infecting 200 ml of exponentially growing *E. faecalis* at a multiplicity of infection of 0.1 in BHI broth containing 10 mM $MgSO_4$. Lysis was allowed to proceed for up to six hours at 37 °C with shaking. The lysates were centrifuged at 10,000*g* for 20 min at room temperature to remove the remaining bacterial cells and debris. Supernatant was then vacuum-filtered through a 0.2-µm membrane filter and kept at 4 °C until use.

Before mice were gavaged, 10–20 ml lysates were concentrated using Corning Spin-X UF Concentrators with 100,000-molecular weight cutoff (MWCO) to a volume of approximately 1 ml. Following concentration, the culture medium was replaced with PBS via diafiltration. The resulting lysate was further concentrated to a final volume of 0.5 ml and adjusted to the required PFUs.

Whole-genome sequencing for phages

For all phages except Efmus4, 10 ml of lysates were treated with 10 µg/ml each of DNase and RNase at 37 °C for 1 h and phages were precipitated by adding 1M NaCl and 10% (w/v) polyethylene glycol 8000 (PEG 8000) and incubated at 4 °C overnight. Precipitated phages were then pelleted by centrifugation at 10,000*g* for 10 min at 4 °C and resuspended in 500 µl of resuspension buffer (5 mM $MgSO_4$). Phage DNA was then extracted using Promega Wizard DNA Clean-up kit (Promega). Phage genomes were sequenced using a combination of Illumina and Oxford Nanopore Technologies (ONT) MinION platforms. Illumina sequencing libraries were prepared using the Nextera XT library kit with bead-based size selection before loading onto Illumina flow cells. Sequencing was performed with either Illumina MiSeq Reagent Kit v3 in 2 × 300-bp or NextSeq 500 Mid Output Kit in 2 × 150-bp paired-end formats. ONT MinION sequencing libraries were prepared using the Rapid Barcoding Kit (SQK-RBK004) and loaded onto MinION R9.4 flow cells. ONT reads were basecalled with Albacore v.2.3.4 (ONT). The sequence reads were demultiplexed and adapters trimmed from ONT reads using Porechop v.0.2.3[36]. A hybrid Illumina-ONT de novo assembly was performed using the Unicycler v.0.4.7 pipeline[37]. Subsequently, Pilon v.1.22[38] was used iteratively to polish the assemblies with Illumina reads until no additional corrections could be made.

For phage Efmus4, 10^9 PFUs of the phage was filtered sequentially using 0.45-µm and 0.2-µm filters (VWR) and purified on a caesium chloride (CsCl) density gradient[39]. One millilitre of the CsCl fraction was purified on Amicon YM-100 protein columns (Millipore) and treated with DNase I. DNA was isolated using a Qiagen UltraSens virus kit (Qiagen), amplified using GenomiPhi V2 (GE Healthcare), and fragmented to 200 to 400 bp using a Bioruptor (Diagenode). Libraries were created using the Ion Plus fragment library kit and sequenced using a 316 Chip on an Ion Torrent Personal Genome Machine (Life Technologies). Reads were trimmed according to modified Phred scores of 0.5 using CLC Genomics Workbench 4.9 (Cambridge), and the remaining reads were assembled using CLC Genomics Workbench 4.9 based on 98% identity with a minimum of 50% read overlap[39]. Reads were assembled into a single contig of 18,186 bp (20,118 × coverage).

Mapping of ONT reads to the hybrid assemblies was used to determine the orientation and terminal ends of linear phage genomes, and reference genomes served as guides to orient circular phage genomes. Phage genome assemblies were annotated using the NCBI Prokaryotic Genome Annotation Pipeline (PGAP)[40,41].

Phage raw sequence reads and annotated genomes are available at NCBI under the following consecutive BioSample IDs (SAMN11089809–SAMN11089827). GenBank accession numbers include: Efmus1 (MK721195), Efmus2 (MK721197), Efmus3 (MK721185), Efmus4 (MK721193), Ef2.1 (MK693030), Ef2.2 (MK721189), Ef2.3 (MK721192), Ef5.1 (MK721199), Ef5.2 (MK721186), Ef5.3 (MK721200), Ef5.4 (MK721191), Ef6.1 (MK721187), Ef6.2 (MK721188), Ef6.3 (MK721196), Ef6.4 (MK721190), Ef7.1 (MK721194), Ef7.2 (MK721183), Ef7.3 (MK721184) and Ef7.4 (MK721198).

Genetic maps of phage genomes were generated by LinearDisplay.pl (https://github.com/JCVenterInstitute/LinearDisplay), a PERL script that uses Xfig (https://sourceforge.net/projects/mcj/) to render high-quality images. Preliminary annotation of genes was derived from the automated annotation and from Phage_Finder[42], which uses curated hidden Markov models and databases of core phage gene to annotate core gene functions. Annotation was then manually reviewed to assign the final colours.

Phage phylogenetic tree

A phage whole-genome phylogeny tree was generated from a pairwise distance matrix calculated with the MASH program, which approximates average nucleotide identity (ANI)[43]. First, a sketch file was created from all the 19 *E. faecalis* phage genomes isolated and sequenced in this study plus 54 *Enterococcus* phage genomes obtained from GenBank, with 5,000 12-mers generated per genome (mash sketch -k 12 -s 5000). The sketch file was then compared to all the initial phage genome sequences to generate the ANI matrix using the mash distance command using default settings. The GGRaSP R package was used to calculate the UPMGA phylogeny from the ANI distance matrix, after redundant phage genomes (genomes ANI >99.985) were removed using the GGRaSP R package with a user defined cutoff of 0.015 (ggrasp.cluster (threshold = 0.015)). The resulting dendrogram was translated into newick format using the APE R package[44], loaded into the iTOL tree viewer[45], and annotated with taxonomic information and manually entered clade identification.

Electron microscopy

Phage morphology was examined by transmission electron microscopy of negatively stained grids, prepared using the valentine method[46] with either 2% uranyl-acetate or 2% phosphotungstic acid, and examined at an acceleration voltage of 100 kV in the JEOL 1200 EX transmission electron microscope.

Bacterial DNA extraction and 16S rRNA sequencing

DNA from human stool samples, mouse liver sections or bacterial culture was extracted as previously described[11], and DNA from mouse faeces was extracted using QIAamp Fast DNA Stool kit (Qiagen). 16S rRNA PCR was completed using Illumina adaptor and barcode-ligated 16S primers targeting the V4 region of the 16S rRNA gene[47,48]. Amplicons were purified using the Qiaquick PCR purification kit (Qiagen) using manufacturer's specifications. Purified amplicons were then quantified via TECAN assay (Tecan), normalized and pooled in preparation for 16S rRNA sequencing. Pooled library was quantified and checked for quality using Agilent 2100 Bioanalyzer (Agilent Technologies). Library was sequenced on Illumina MiSeq (Illumina) using V2 reagent chemistry, 500 cycles, 2 × 250-bp format using manufacturer's specifications. 16S sequence reads were processed and operational taxonomic units were determined using our MOTHUR-based 16S rDNA analysis workflow as previously described[11,49]. Raw 16S sequence reads can be found in the NCBI Sequence Read Archive (SRA) associated with Bioproject PRJNA525701.

Real-time qPCR

Bacterial genomic DNA was extracted from human stool samples and mouse liver[11]. RNA was extracted from mouse liver and cDNAs were generated[11]. Primer sequences for mouse genes were obtained from the NIH qPrimerDepot. Primer sequences for *E. faecalis* 16S rRNA gene, *E. faecalis* $cylL_S$ and $cylL_L$ genes have previously been described[50,51]. All primers used in this study are listed in Supplementary Table 9. Mouse gene expression and amplification of bacterial genes were determined with Sybr Green (Bio-Rad Laboratories) using ABI StepOnePlus real-time PCR system. The qPCR value of mouse genes was normalized to 18S.

E. faecalis isolation and whole-genome sequencing

To isolate *E. faecalis* strains from human subjects, 50–300 mg of human stool was resuspended in 500 µl PBS, serial dilutions were made and 100 µl was placed on plates with selective medium, BBL Enterococcosel broth (Becton Dickinson). *Enterococci* colonies were identified by the production of dark brown or black colour, generated by hydrolysis of esculin to esculetin (which reacts with ferric ammonium citrate). Each *Enterococcus* colony was then picked, and qPCR was performed to identify *E. faecalis*, using specific primers against the *E. faecalis* 16S rRNA gene[50]. For each subject, between 1 and 6 *E. faecalis* colonies were analysed and bacterial genomic DNA was then extracted as described in 'Bacterial DNA extraction and 16S rRNA sequencing'.

DNA sequencing was performed on the Illumina HiSeq Ten X generating paired-end reads (2 × 151 bp). Bacterial genomes were assembled

and annotated using the previously described pipeline[52]. Antimicrobial resistance and virulence genes including cytolysin (*cyl*) genes carried by *E. faecalis* isolates were identified by comparing individual genome assemblies against the CARD and VFDB databases, respectively, using abricate v0.8.10 (https://github.com/tseemann/abricate)[53,54].

For the phylogeny of *E. faecalis*, the genome assemblies of the study isolates were annotated with Prokka[55], and a pangenome estimated using Roary[52]. A 95% identity cutoff was used, and core genes were defined as those in 99% of isolates. A maximum likelihood tree of the SNPs in the core genes was created using RAxML[56] and 100 bootstraps. The resulting tree was visualized using iTOL[45]. Genome sequence data of *E. faecalis* strains isolated in this study have been deposited in the European Nucleotide Archive (ENA) under the accession number PRJEB25007. Sequence reads are available at ENA under run accession identifiers ERR3200171–ERR3200263.

E. faecalis culture

All *E. faecalis* strains were grown statically in BHI broth or on BHI agar plate at 37 °C. Fifty micrograms per millilitre erythromycin was added when cytolytic and non-cytolytic *E. faecalis* strains were grown (Extended Data Fig. 2).

Determination of levels of faecal *Enterococcus*

To determine levels of faecal enterococci in mice, 10–30 mg of mouse faeces was resuspended into 500 µl PBS and serial dilutions were made. Five microlitres of each dilution from each sample was spotted onto a plate with a selective medium, BBL enterococcosel broth (Becton Dickinson) and the plates were then incubated at 37 °C overnight. For Extended Data Fig. 2o, agar plates contained 50 µg/ml erythromycin. Enterococci colonies were identified by the production of a dark brown or black colour. Colony numbers of each sample were then counted, and CFUs were calculated.

Cytolysin expression and purification

To purify bioactive $CylL_L''$ and $CylL_S''$, an *Escherichia coli* heterologous expression system was used. In brief, either 6×His–$CylL_L$ or 6×His–$CylL_S$ were co-expressed with CylM (the enzyme that performs dehydration and cyclization reactions on cytolysin) in *E. coli* to yield fully dehydrated and cyclized full-length peptides. The His tag and leader peptide were then removed using recombinant CylA (27–412), the soluble domain of the native peptidase used in cytolysin maturation, to yield bioactive $CylL_L''$ or $CylL_S''$. The resulting core peptides were further purified by reversed-phase high-performance liquid chromatography (HPLC).

The $cylL_L$ and $cylL_S$ genes were previously cloned into the MCSI of a pRSFDuet-1 backbone vector that contained the *cylM* gene in MCSII[6]. The *cylA* (27–412) gene was previously cloned into MCSI of a pRSFDuet-1 backbone vector[57]. *E. coli* BL21 Star (DE3) cells (50 µl) were transformed with 100 ng of either the $cylL_L$_*cylM*:pRSFDuet, $cylL_S$_*cylM*:pRSFDuet or *cylA* (27–412):pRSFDuet plasmids via KCM chemical transformation. The cells were plated on LB agar plates supplemented with kanamycin (50 µg/ml) and grown at 37 °C overnight. One colony was picked to inoculate 15 ml of LB broth supplemented with kanamycin overnight at 37 °C. The culture was used to inoculate 1.5 l of terrific broth supplemented with kanamycin. Cultures were grown with shaking at 37 °C to an optical density at 600 nm (OD_{600}) of 0.8. The temperature of the incubator was lowered to 18 °C and expression was induced with the addition of 0.3 mM final concentration of isopropyl β-D-thiogalactoside. The cultures were allowed to incubate at 18 °C for 18 h. The cells were collected by centrifugation at 5,000*g* for 12 min. The cell paste was collected and frozen at -70 °C.

For the purification of the protease CylA (27–412), the cell paste was thawed and resuspended in 50 ml LanP buffer (20 mM HEPES, 1 M NaCl, pH 7.5). The cell suspension was lysed by homogenization. The lysate was clarified by centrifugation at 13,000*g* for 45 min and filtered through a 0.45-µm centrifugal filter (Thermo Scientific). The clarified lysate was applied to a pre-equilibrated HisTrap HP 5 ml column (GE Healthcare) through a peristaltic pump. The loaded column was connected to an ÄKTA pure 25 M system. The protein was eluted by a linear gradient of LanP buffer to Elution Buffer (20 mM HEPES, 1 M NaCl, 500 mM imidazole, 10% glycerol, pH 7.5) over 30 min. The purest fractions, as determined by 4–20% SDS–PAGE, were combined, concentrated to 1 mg/ml by Amicon Ultra Centrifugal Filters (30 kDa MWCO), and buffer exchanged into storage buffer (20mM HEPES, 300 mM KCl, 10% glycerol, pH 7.5) by PD-10 desalting column (GE Healthcare). Protein concentration was determined by absorbance at 280 nm.

For the purification of $CylL_L''$ and $CylL_S''$ peptides, the cell paste was thawed and resuspended in 50 ml of LanA Buffer B1 (6 M guanidine HCl, 20 mM NaH_2PO_4, 500 mM NaCl, 0.5 mM imidazole, pH 7.5). The cell suspension was lysed via sonication (2-s pulse on, 5-s pulse off, 7 min total pulse on time). The cell lysate was clarified by centrifugation at 13,000*g* for 45 min. The clarified cell lysate was filtered through a 0.45-µm centrifugal filter and applied via gravity flow to a pre-equilibrated, 2 ml bed volume of His60 Ni Superflow Resin (Clonetech). After the lysate had been applied, the resin was washed with 15 ml of LanA Buffer B2 (4 M guanidine HCl, 20 mM NaH_2PO_4, 500 mM NaCl, 30 mM imidazole, pH 7.5). The resin was washed again with 15 ml of LanA Wash Buffer (20 mM NaH_2PO_4, 500 mM NaCl, 30 mM imidazole, pH 7.5) to remove the guanidine HCl. The peptides were eluted with 10 ml of LanA elution buffer (20 mM NaH_2PO_4, 500 mM NaCl, 500 mM imidazole, pH 7.5). A 0.02 mg/ml final concentration of CylA (27–412) was added to the elution fraction and allowed to incubate at room temperature overnight to remove the leader peptide.

The digestion was quenched by adding 2% (v/v) final concentration of trifluoroacetic acid. The solution was centrifuged at 4,500*g* for 10 min and filtered through a 0.45-µm syringe filter (Thermo Scientific). The core peptides were purified by semi-preparative reverse-phase HPLC using a Phenomenex Jupiter Proteo column (10 mm × 250 mm, 4 µm, 90 Å) connected to an Agilent 1260 Infinity II liquid chromatography system. The peptides were separated using a linear gradient of 3% (v/v) solvent B (acetonitrile + 0.1% trifluoroacetic acid) in solvent A (water + 0.1% trifluoroacetic acid). The fractions were spotted on a matrix-assisted laser desorption/ionization (MALDI) target plate by mixing 1 µl of sample with 1 µl of a 25 mg/ml solution of Super-DHB (Sigma) in 80% acetonitrile/water + 0.1% trifluoroacetic acid. The fractions were analysed by MALDI–time of flight (TOF) mass spectrometry on a Bruker UltrafleXtreme MALDI–TOF/TOF operating in positive ionization, reflector mode.

Primary mouse hepatocytes

Hepatocytes were isolated from C57BL/6 female mice fed the chronic–binge ethanol diet (NIAAA model)[15]. Livers were perfused in situ with 0.5 mM EGTA containing calcium-free salt solution and then perfused with a solution containing 0.02% (w/v) collagenase D (Roche Applied Science). Livers were then carefully minced and filtered using a 70-µm nylon cell strainer. Hepatocytes were centrifuged at 50*g* for 1 min after 3 washes. Hepatocyte viability was assessed by Trypan Blue (Thermo Fisher Scientific). Hepatocytes (1.5×10^5) were seeded on 12-well plates coated with rat collagen type I in DMEM-F12 (Thermo Fisher Scientific) with insulin–transferrin–selenium (1% v/v) (Thermo Fisher Scientific) and 40 ng/ml dexamethasone (MP Biomedicals) containing 10% (v/v) fetal bovine serum (FBS) (Gemini Bio-Products) and antibiotics. After 4 h, the culture was washed with DMEM-F12 medium and changed to the same complemented medium without FBS[58]. Then 16 h later, hepatocytes were cultured with 0 or 25 mM ethanol and stimulated with 0, 200 or 400 nM $CylL_S''$ and/or $CylL_L''$ in the same culture medium without FBS. After 3 h stimulation, hepatocyte cytotoxicity was assessed using Pierce LDH cytotoxicity detection kit (Thermo Fisher Scientific).

Biochemical analysis

Serum levels of ALT were determined using Infinity ALT kit (Thermo Scientific). Hepatic triglyceride levels were measured using Triglyceride

Liquid Reagents kit (Pointe Scientific). Levels of serum lipopolysaccharide and faecal albumin were determined by enzyme-linked immunosorbent kits (Lifeome Biolabs and Bethyl Labs, respectively). Serum levels of ethanol were measured using ethanol assay kit (BioVision).

Staining procedures

Formalin-fixed tissue samples were embedded in paraffin and stained with H & E. To determine lipid accumulation, liver sections were embedded in OCT compound. Eight-micrometre frozen sections were then cut and stained with Oil Red O (Sigma-Aldrich). Representative images from each group of mice are shown in each figure. The terminal deoxynucleotide transferase-mediated dUTP nick-end labelling (TUNEL) assay was performed using an in situ cell death detection kit (Sigma-Aldrich). We randomly selected five high-power fields for counting TUNEL-positive cells and normalized numbers to total cells.

Statistical analysis

Results are expressed as mean ± s.e.m. (except when stated otherwise). Univariate and multivariate Cox regression analysis was used to detect associations of cytolysin with overall mortality. The multivariate model was adjusted for geographical origin of the patients, antibiotic treatment, platelet count, and creatinine, bilirubin and INR as components of the MELD score. Univariate logistic regression analysis of laboratory and clinical parameters associated with the detection of cytolysin and *E. faecalis* was performed. Univariate linear regression analysis of laboratory and clinical parameters associated with the log-transformed total amount of faecal *E. faecalis* measured with qPCR was performed. To associate log-transformed total *E. faecalis* and *E. faecalis* positivity with mortality, univariate Cox regression was used. *P* values from univariate and multivariate Cox regression, univariate logistic regression and univariate linear regression were determined by Wald test. Multicollinearity was examined using the variance inflation factor. Kaplan–Meier curves were used to compare survival between cytolysin-positive and cytolysin-negative patients with alcoholic hepatitis. Faecal *E. faecalis*, bacterial diversity and richness from controls and patients were compared using Kruskal–Wallis test with Dunn's post hoc test. Region- and/or centre-specific clinical characteristics of patients with alcoholic hepatitis were compared with Kruskal–Wallis test for continuous and Fisher's exact test for categorical variables. Faecal *E. faecalis* in patients with alcoholic hepatitis with or without cytolysin, and with or without cirrhosis, were compared with Mann–Whitney-Wilcoxon rank-sum test. Faecal *E. faecalis* in patients with alcoholic hepatitis from different region and/or centres were compared with the Kruskal–Wallis test. The percentage of subjects with faecal samples that were positive for *E. faecalis* and cytolysin was compared using Fisher's exact test, followed by FDR procedures for multiple group comparisons. Jaccard dissimilarity matrices were used for PCoA, and *P* values were determined by PERMANOVA followed by FDR procedures to correct for multiple comparisons.

For mouse and cell culture studies, the significance of multiple groups was evaluated using one-way or two-way ANOVA with Tukey's post hoc test. Fisher's exact test was used in the analysis of liver *E. faecalis* and cytolysin with FDR correction for multiple comparisons. Kaplan–Meier curves were used to compare survival between experimental mouse groups. PCoA based on Jaccard dissimilarity matrices was performed between experimental mouse groups and the *P* values were determined by PERMANOVA followed by FDR procedures to correct for multiple comparisons.

Exact *P* values for all comparisons, together with group size for each group, were listed in Supplementary Table 10. Statistical analyses were performed using R statistical software, R v.3.5.1 (R Foundation for Statistical Computing) and GraphPad Prism v.6.01. A value of $P < 0.05$ was considered to be statistically significant (adjusted for multiple comparisons when performing multiple tests).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Raw 16S sequence reads can be found in the NCBI SRA associated with Bioproject PRJNA525701. Phage raw sequence reads and annotated genomes are available at NCBI under the following consecutive BioSample identifiers (SAMN11089809–SAMN11089827). Genome sequence data of *E. faecalis* strains isolated in this study were registered at the ENA under study PRJEB25007.

Code availability

The PERL script for making the genetic maps of phage genomes can be found at https://github.com/JCVenterInstitute/LinearDisplay.

30. Brandl, K. et al. Dysregulation of serum bile acids and FGF19 in alcoholic hepatitis. *J. Hepatol.* **69**, 396–405 (2018).
31. Gao, B. et al. Serum and fecal oxylipins in patients with alcohol-related liver disease. *Dig. Dis. Sci.* **64**, 1878–1892 (2019).
32. Lang, S. et al. Intestinal fungal dysbiosis and systemic immune response to fungi in patients with alcoholic hepatitis. *Hepatology* (2019).
33. Ball, S. A., Tennen, H., Poling, J. C., Kranzler, H. R. & Rounsaville, B. J. Personality, temperament, and character dimensions and the DSM-IV personality disorders in substance abusers. *J. Abnorm. Psychol.* **106**, 545–553 (1997).
34. Krieg, L. et al. Mutation of the gastric hydrogen-potassium ATPase alpha subunit causes iron-deficiency anemia in mice. *Blood* **118**, 6418–6425 (2011).
35. Gill, J. J. et al. The *Caulobacter crescentus* phage phiCbK: genomics of a canonical phage. *BMC Genomics* **13**, 542 (2012).
36. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb. Genom.* **3**, e000132 (2017).
37. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Comput. Biol.* **13**, e1005595 (2017).
38. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
39. Santiago-Rodriguez, T. M. et al. Transcriptome analysis of bacteriophage communities in periodontal health and disease. *BMC Genomics* **16**, 549 (2015).
40. Haft, D. H. et al. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.* **46**, D851–D860 (2018).
41. Tatusova, T. et al. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44**, 6614–6624 (2016).
42. Fouts, D. E. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res.* **34**, 5839–5851 (2006).
43. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
44. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
45. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
46. Valentine, R. C., Shapiro, B. M. & Stadtman, E. R. Regulation of glutamine synthetase. XII. Electron microscopy of the enzyme from *Escherichia coli*. *Biochemistry* **7**, 2143–2152 (1968).
47. Haas, B. J. et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* **21**, 494–504 (2011).
48. Caporaso, J. G. et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl Acad. Sci. USA* **108** (Suppl 1), 4516–4522 (2011).
49. Chen, P. et al. Supplementation of saturated long-chain fatty acids maintains intestinal eubiosis and reduces ethanol-induced liver injury in mice. *Gastroenterology* **148**, 203–214.e16 (2015).
50. Ryu, H. et al. Development of quantitative PCR assays targeting the 16S rRNA genes of *Enterococcus* spp. and their application to the identification of *enterococcus* species in environmental samples. *Appl. Environ. Microbiol.* **79**, 196–204 (2013).
51. Haas, W., Shepard, B. D. & Gilmore, M. S. Two-component regulator of *Enterococcus faecalis* cytolysin responds to quorum-sensing autoinduction. *Nature* **415**, 84–87 (2002).
52. Page, A. J. et al. Robust high-throughput prokaryote *de novo* assembly and improvement pipeline for Illumina data. *Microb. Genom.* **2**, e000083 (2016).
53. Chen, L. et al. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.* **33**, D325–D328 (2005).
54. Jia, B. et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **45**, D566–D573 (2017)
55. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
56. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

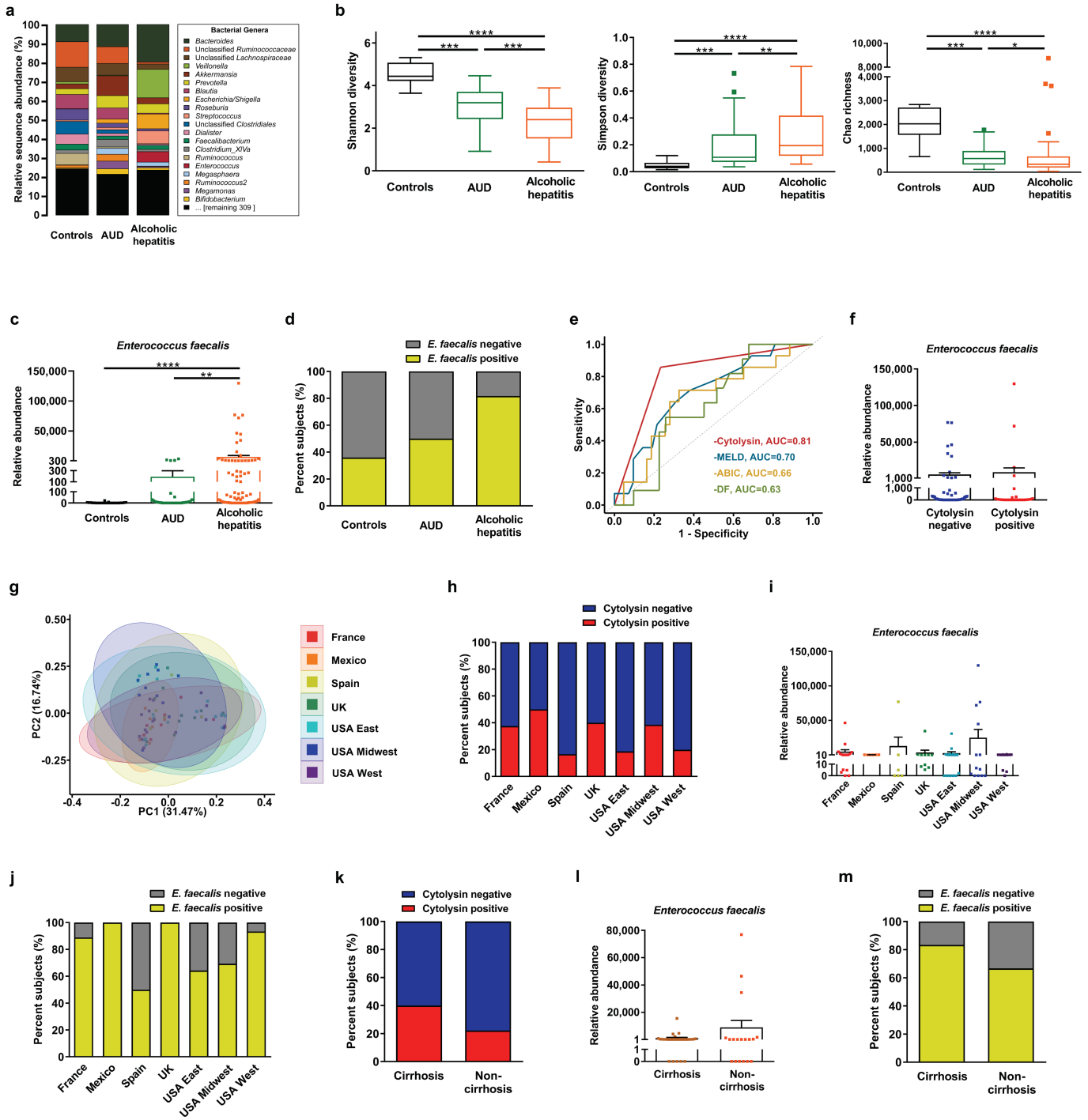
57. Tang, W., Bobeica, S. C., Wang, L. & van der Donk, W. A. CylA is a sequence-specific protease involved in toxin biosynthesis. *J. Ind. Microbiol. Biotechnol.* **46**, 537–549 (2019).
58. Iwaisako, K. et al. Protection from liver fibrosis by a peroxisome proliferator-activated receptor δ agonist. *Proc. Natl Acad. Sci. USA* **109**, E1369–E1376 (2012).
59. Clarke, T. H., Brinkac, L. M., Sutton, G. & Fouts, D. E. GGRaSP: a R-package for selecting representative genomes using Gaussian mixture models. *Bioinformatics* **34**, 3032–3034 (2018).

Acknowledgements S. Lang was supported by a DFG fellowship (LA 4286/1-1), C.L. was supported by an AASLD Pinnacle Research Award in Liver Disease and a pilot project award from Southern California Research Center for Alcoholic Liver and Pancreatic Disease (ALPD) and Cirrhosis (P50 AA011999), Y.S. was supported by a Wellcome Trust PhD studentship and I.R.R. was supported by National Institute of General Medical Sciences (NIGMS)–NIH Chemistry–Biology Interface Training Grant (T32-GM070421). This study was supported in part by a Biocodex Microbiota Foundation Grant, NIH grants R01 AA24726, U01 AA026939, by Award Number BX004594 from the Biomedical Laboratory Research & Development Service of the VA Office of Research and Development (to B.S.), the Wellcome Trust (WT098051) (to T.D.L.), NIH grant U01AA021908 (to R.B.) and services provided by P30 DK120515 and P50 AA011999.

Author contributions Y.D. was responsible for acquisition, analysis and interpretation of data, and drafting of the manuscript; C.L. was responsible for study concept and design, acquisition, analysis and interpretation of data and key preliminary experiments; S. Lang, K.B., J.L. and X.M.T. provided assistance with statistical analysis; H.C., L.J., B.G., W.S., R.K., F.H. and S. Lee provided assistance with data acquisition; R.C.W., T.H.C., K.N., M.T. and D.E.F. were responsible for 16S rRNA sequencing, phage genome sequencing and data analysis; Y.S. and T.D.L. were responsible for bacterial genome sequencing and data analysis; A.H.-M., L.L. and R.Y. provided assistance with phage studies and were responsible for electron microscopy data; R.Y. provided critical revision of the manuscript; I.R.R. and W.A.v.d.D. were responsible for cytolysin expression and purification; Y.M. and L.E. provided assistance with the design and conduct of the gnotobiotic mouse studies; M.L. and D.P. provided assistance with phage isolation; M.V.-C., F.B.-P., E.C.V., J.G.A., R.S.B. Jr, V.V., J.A., J.C., D.L.S., S.B.H., A.L., M.R.L., P.M., G.G.-T., R.B. and P.S. were responsible for collection of human samples; D.E.F. and B.S. were responsible for the study concept and design, and editing the manuscript; B.S. was responsible for study supervision.

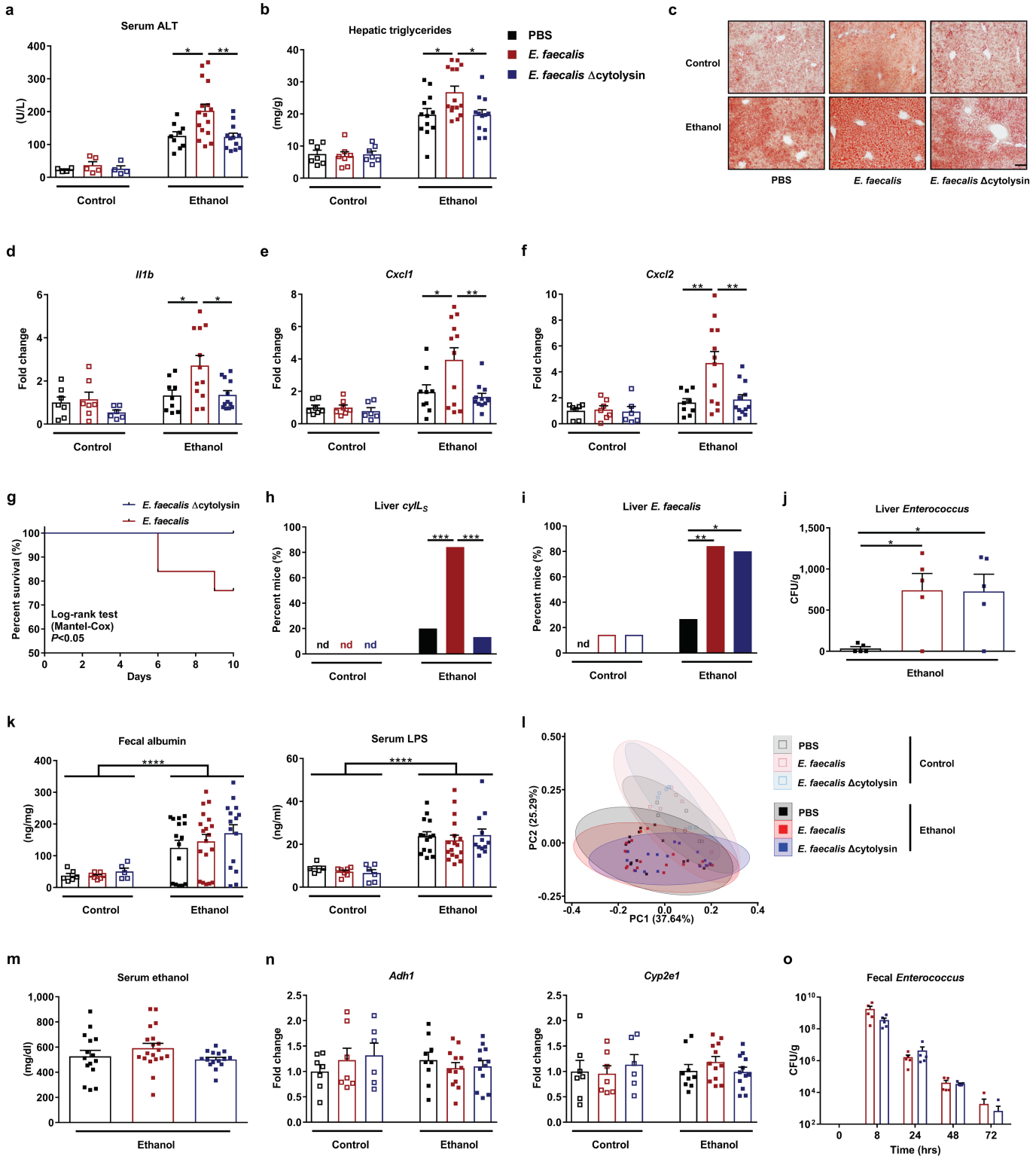
Competing interests B.S. consults for the Ferring Research Institute; however, there is no competing interest with regard to this study. All other authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1742-x.
Correspondence and requests for materials should be addressed to B.S.
Reprints and permissions information is available at http://www.nature.com/reprints.



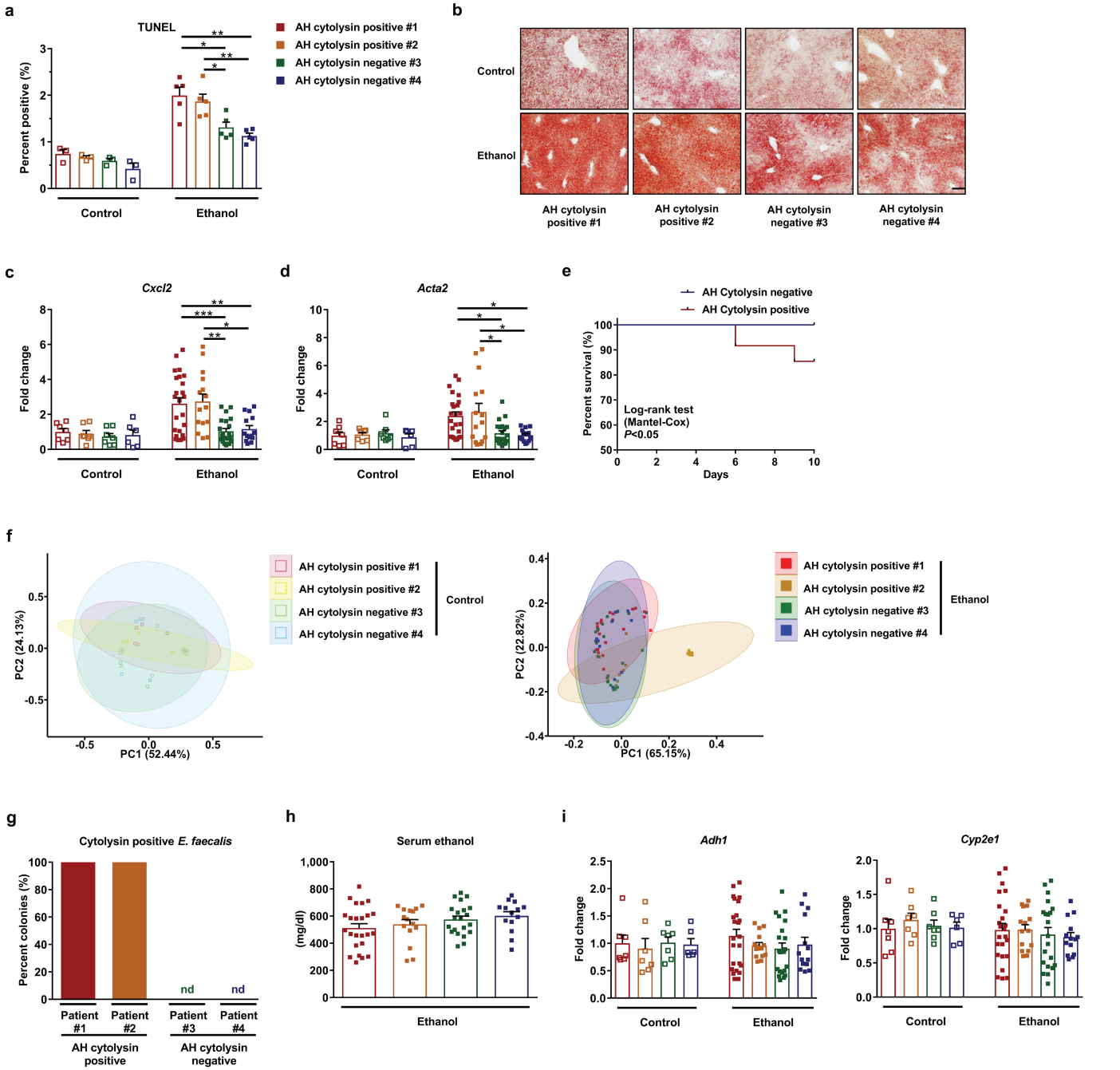
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Intestinal dysbiosis in patients with alcoholic hepatitis. a, 16S rRNA sequencing of faecal samples from controls ($n = 14$), patients with alcohol-use disorder ($n = 43$), or alcoholic hepatitis ($n = 75$). The graph demonstrates the relative abundance of sequence reads in each genus. **b**, Bacterial diversity (Shannon index and Simpson index) and richness (Chao richness) was calculated in controls ($n = 14$), patients with alcohol-use disorder ($n = 43$) or alcoholic hepatitis ($n = 75$). **c**, *E. faecalis* in faecal samples from controls ($n = 25$), patients with alcohol-use disorder ($n = 38$) or alcoholic hepatitis ($n = 82$), assessed by qPCR. **d**, Percentage of faecal samples positive for *E. faecalis* in controls ($n = 25$), patients with alcohol-use disorder ($n = 38$) or alcoholic hepatitis ($n = 82$), assessed by qPCR. *E. faecalis* was detected in faeces from 80% of patients with alcoholic hepatitis, versus 36% of controls ($P < 0.001$). There was also a significant difference between patients with alcohol-use disorder and patients with alcoholic hepatitis ($P < 0.01$). **e**, Receiver operating characteristic curves and area under the curve (AUC) for the comparison of 90-day mortality and cytolysin positivity (red; $n = 57$), MELD score (blue; $n = 56$), ABIC score (yellow; $n = 57$) and discriminant function (green; $n = 42$) in patients with alcoholic hepatitis. **f**, *E. faecalis* in faecal samples from patients with alcoholic hepatitis whose faecal samples were cytolysin-positive ($n = 25$) or cytolysin-negative ($n = 54$), assessed by qPCR ($P = 0.8174$). **g**, 16S rRNA sequencing of faecal samples from patients with alcoholic hepatitis from different centres (France, $n = 9$; Mexico, $n = 6$; Spain, $n = 5$; UK, $n = 11$; USA (east), $n = 16$; USA (Midwest), $n = 12$; USA (west), $n = 16$ patients). We used PCoA based on Jaccard dissimilarity matrices to show β-diversity among groups at the genus level. The composition of faecal microbiota was significantly different between patients from different regions ($P < 0.01$). **h**, Percentage of faecal samples that were positive for $cylL_L$ and $cylL_S$ DNA sequences (cytolysin-positive), in patients with alcoholic hepatitis from different centres (France, $n = 16$; Mexico, $n = 6$; Spain, $n = 6$; UK, $n = 10$; USA (east), $n = 16$; USA (Midwest), $n = 13$; USA (west), $n = 15$ patients), assessed by qPCR ($P = 0.6094$). **i**, *E. faecalis* in faecal samples from patients with alcoholic hepatitis from different centres, assessed by qPCR ($P = 0.5648$). **j**, Percentage of faecal samples that were positive for *E. faecalis* in patients with alcoholic hepatitis from different centres (France, $n = 16$; Mexico, $n = 6$; Spain, $n = 6$; UK, $n = 10$; USA (east), $n = 16$; USA (Midwest), $n = 13$; USA (west), $n = 15$ patients), assessed by qPCR ($P = 0.0529$). **k**, Percentage of subjects with faecal samples that were positive for $cylL_L$ and $cylL_S$ DNA sequences (cytolysin-positive), in patients with alcoholic hepatitis and with ($n = 30$) or without ($n = 18$) cirrhosis, assessed by qPCR ($P = 0.3431$). **l**, *E. faecalis* in faecal samples from patients with alcoholic hepatitis and with ($n = 30$) or without ($n = 18$) cirrhosis, assessed by qPCR ($P = 0.5736$). **m**, Percentage of faecal samples that were positive for *E. faecalis* in patients with alcoholic hepatitis and with ($n = 30$) or without ($n = 18$) cirrhosis, assessed by qPCR ($P = 0.2878$). Results are expressed as mean ± s.e.m. (**c**, **f**, **i**, **l**). For the box and whisker plots in **b**, the box extends from the 25th to 75th percentiles, and the centre line represents the median; for all three groups, the bottom whiskers show the minimum values; for the control group (black), the top whisker shows the maximum value; for the other two groups, the top whiskers represent the 75th percentile plus 1.5× the inter-quartile distance (the distance between the 25th and 75th percentiles); all values greater than this are plotted as individual dots. *P* values were determined by Kruskal–Wallis test (**i**) with Dunn's post hoc test (**b**, **c**), two-sided Fisher's exact test (**h**, **j**, **k**, **m**) followed by FDR procedures (**d**), two-sided Mann–Whitney Wilcoxon rank-sum test (**f**, **l**) or PERMANOVA (**g**). The exact group size (n) and P values for each comparison are listed in Supplementary Table 10. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

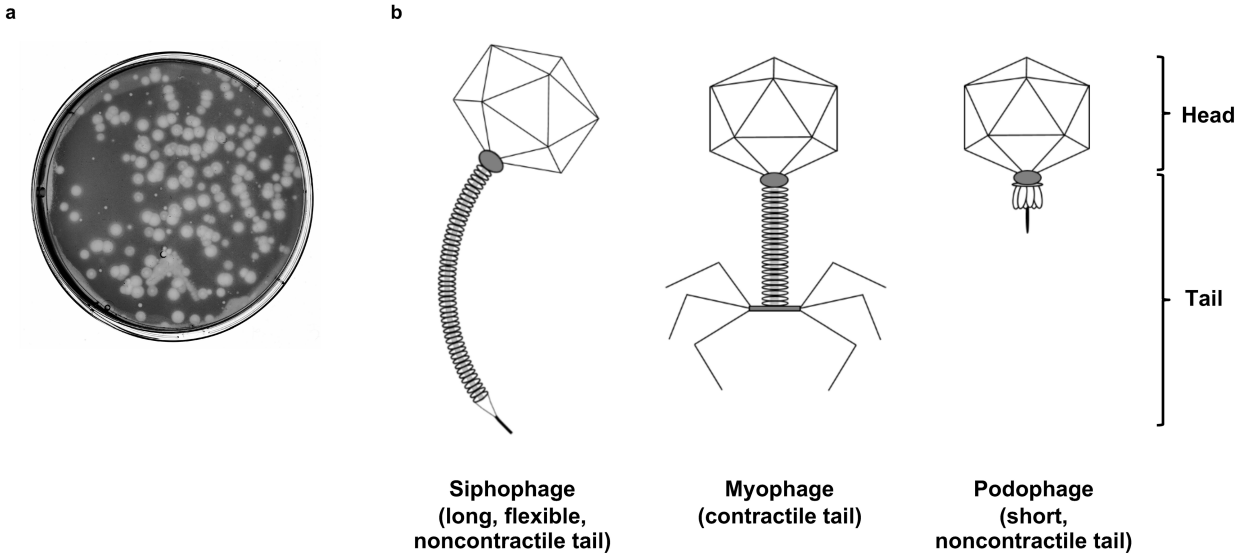


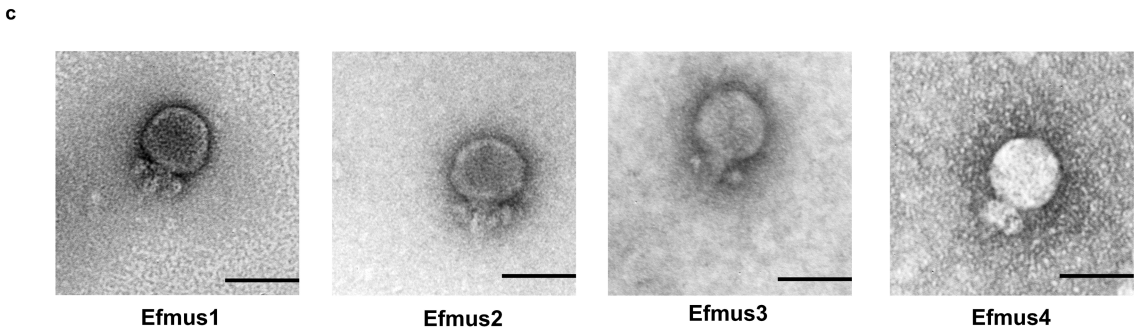
Extended Data Fig. 2 | See next page for caption.

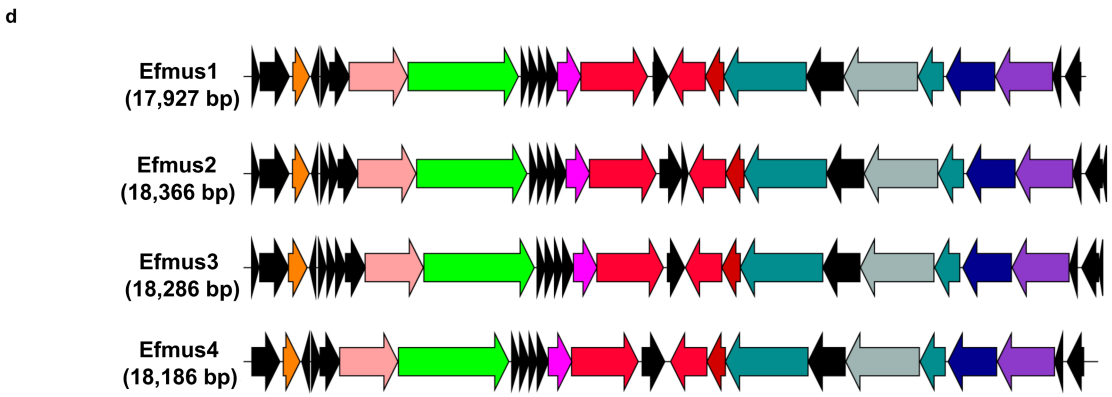
Extended Data Fig. 2 | Cytolytic *E. faecalis* causes the progression of ethanol-induced liver disease in mice. a–**n**, C57BL/6 mice were fed oral isocaloric (control) or chronic–binge ethanol diets and gavaged with vehicle (PBS), a cytolytic *E. faecalis* strain (FA2-2(pAM714)) (denoted *E. faecalis*) (5×10^8 CFUs) or a non-cytolytic *E. faecalis* strain (FA2-2(pAM771))[5] (denoted *E. faecalis* Δcytolysin) (5×10^8 CFUs) every third day. **a**, Serum levels of ALT. **b**, Hepatic triglyceride content. **c**, Representative oil red O-stained liver sections. **d**–**f**, Hepatic levels of mRNAs. **g**, Kaplan–Meier curve of survival of mice on chronic–binge ethanol diets (day 0 denotes the start of ethanol feeding). Mice gavaged with PBS all survived, and are not included in the figure. A higher proportion of mice ($n = 15$) gavaged with non-cytolytic *E. faecalis* survived than did mice ($n = 25$) gavaged with cytolytic *E. faecalis*. **h**, Proportions of mice that were positive for cytolysin in the liver, measured by qPCR for $cylL_S$ (the gene that encodes cytolysin subunit $CylL_S''$). **i**, Proportions of mice that were positive for *E. faecalis* in the liver, measured by qPCR. About 80% of mice colonized with cytolytic *E. faecalis*, as well as those colonized with non-cytolytic *E. faecalis*, were positive for *E. faecalis* in their livers. **j**, Liver CFUs of *Enterococcus* in mice on a chronic–binge ethanol diet. **k**, Paracellular intestinal permeability was evaluated by measuring faecal albumin content and serum levels of lipopolysaccharide (LPS) by enzyme-linked immunosorbent assays. **l**, Faecal samples were collected and 16S rRNA genes were sequenced. PCoA based on Jaccard dissimilarity matrices showed no significant differences among mice gavaged with PBS, cytolytic or non-cytolytic *E. faecalis* following feeding with the control and ethanol diets. Compared to mice fed with a control diet, mice fed with an ethanol diet had significantly different faecal microbiomes after gavaging with *E. faecalis* ($P < 0.05$). **m**, **n**, Serum levels of ethanol and hepatic levels of *Adh1* and *Cyp2e1* mRNAs did not differ significantly among mice gavaged with PBS, cytolytic or non-cytolytic *E. faecalis* after ethanol feeding. **o**, Mice were gavaged with cytolytic or non-cytolytic *E. faecalis* strains (carrying the erythromycin resistance gene; 5×10^8 CFUs) at time 0, and faeces were collected 0, 8, 24, 48 and 72 h later. Faecal CFUs of *Enterococcus* were determined by culturing faecal samples on BBL enterococcosel broth agar plate with 50 μg ml^{-1} erythromycin. At time 0 and 72 h, five out of five and four out of five mice, respectively, had no detectable erythromycin-resistant *Enterococcus* in their faeces. These points are not shown on the graph, but have been included in the calculation of mean ± s.e.m. Scale bar, 100 μm. Results are expressed as mean ± s.e.m. (**a**, **b**, **d**–**f**, **j**, **k**, **m**–**o**). *P* values among groups of mice fed with the control or ethanol diet were determined by one-way ANOVA with Tukey's post hoc test (**a**, **b**, **d**–**f**, **j**, **k**, **m**, **n**), two-sided log-rank (Mantel–Cox) test (**g**), two-sided Fisher's exact test followed by FDR procedures (**h**, **i**) or PERMANOVA followed by FDR procedures (**l**). All results were generated from at least three independent replicates. The exact group size (n) and *P* values for each comparison are listed in Supplementary Table 10. *P* values between mice fed with a control diet and mice fed with an ethanol diet were determined by two-way ANOVA (**k**). $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.



Extended Data Fig. 3 | Transplantation of cytolysin-positive faeces increases ethanol-induced liver disease in gnotobiotic mice. a–**f**, **h**, **i**, C57BL/6 germ-free mice were colonized with faeces from two cytolysin-positive and two cytolysin-negative patients with alcoholic hepatitis, and then fed isocaloric (control) or chronic–binge ethanol diets. **a**, Percentage of TUNEL-positive hepatic cells. **b**, Representative oil red O-stained liver sections. **c**, **d**, Hepatic levels of mRNAs that encode the inflammatory cytokine *Cxcl2* and *Acta2* (a marker of activated hepatic stellate cells). **e**, Kaplan–Meier curve of survival of mice on chronic–binge ethanol diets (day 0 denotes the start of ethanol feeding), gavaged with faeces from cytolysin-positive ($n = 48$ mice) or cytolysin-negative ($n = 32$ mice) patients with alcoholic hepatitis. **f**, Faecal samples were collected and 16S rRNA genes were sequenced. The graph shows PCoA of faecal microbiomes. No significant difference was observed between mice colonized with faeces from cytolysin-positive or cytolysin-negative donors with alcoholic hepatitis, following the control diet. Mice transplanted with faeces from a cytolysin-positive patient with alcoholic hepatitis (patient no. 2) showed a microbiota that was significantly different to that of the other mouse groups following ethanol administration ($P < 0.01$). **g**, Percentage of cytolysin-positive *E. faecalis* in four patients with alcoholic hepatitis. Stool samples from the four patients were placed on plates with selective medium, and *Enterococcus* colonies were identified by the production of a dark brown or black colour. *Enterococcus* colonies were confirmed to be *E. faecalis* by qPCR. The cytolysin status of each *E. faecalis* colony was determined by qPCR. **h**, Serum levels of ethanol were comparable among colonized mice after ethanol feeding. **i**, Hepatic levels of *Adh1* and *Cyp2e1* mRNAs did not differ significantly among colonized mice on control or ethanol diets. Scale bar, 100 µm. Results are expressed as mean ± s.e.m. (**a**, **c**, **d**, **h**, **i**). *P* values were determined by one-way ANOVA with Tukey's post hoc test (**a**, **c**, **d**, **h**, **i**), two-sided log-rank (Mantel–Cox) test (**e**) or PERMANOVA followed by FDR procedures (**f**). All results were generated from at least three independent replicates. The exact group size (n) and *P* values for each comparison are listed in Supplementary Table 10. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

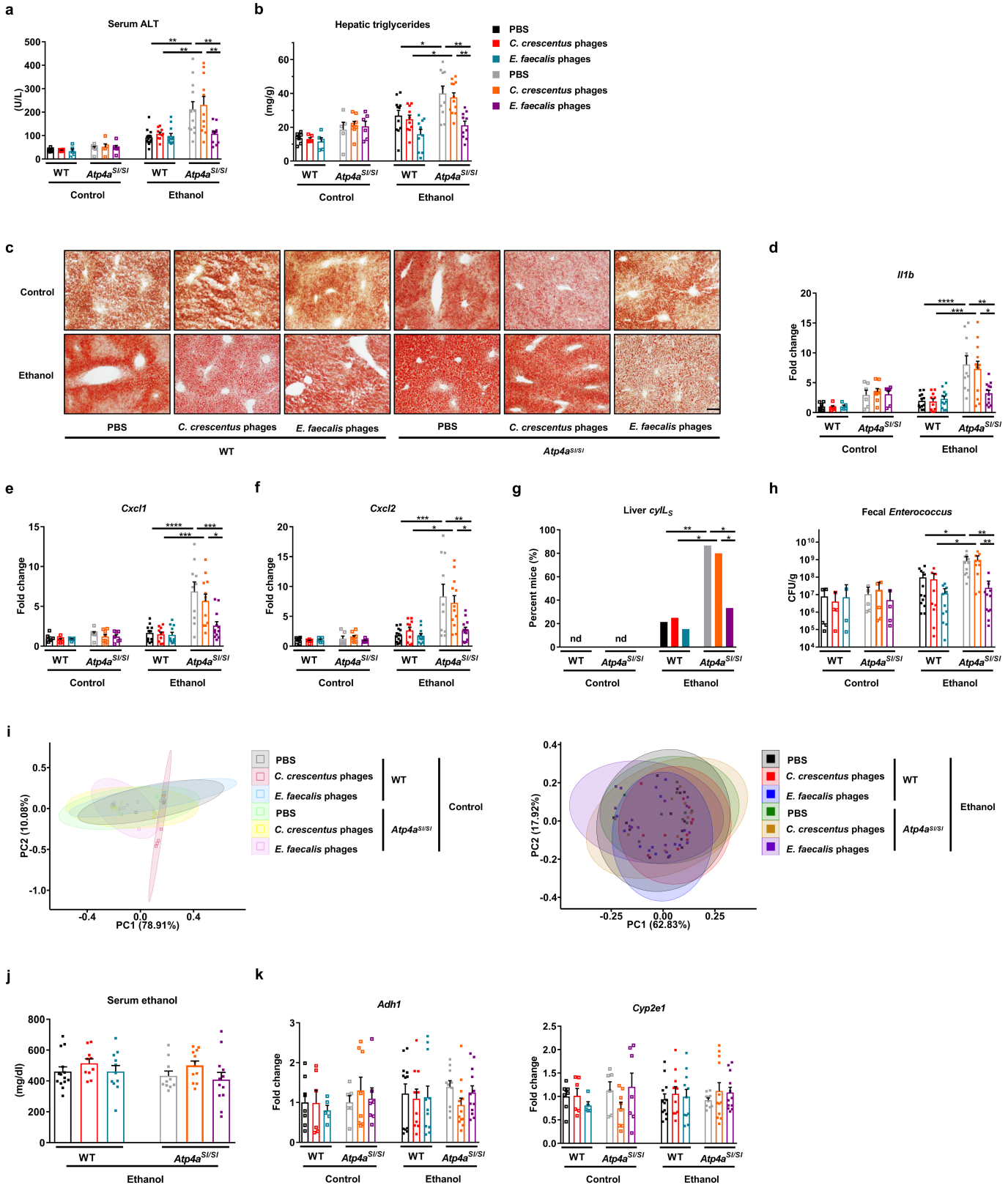






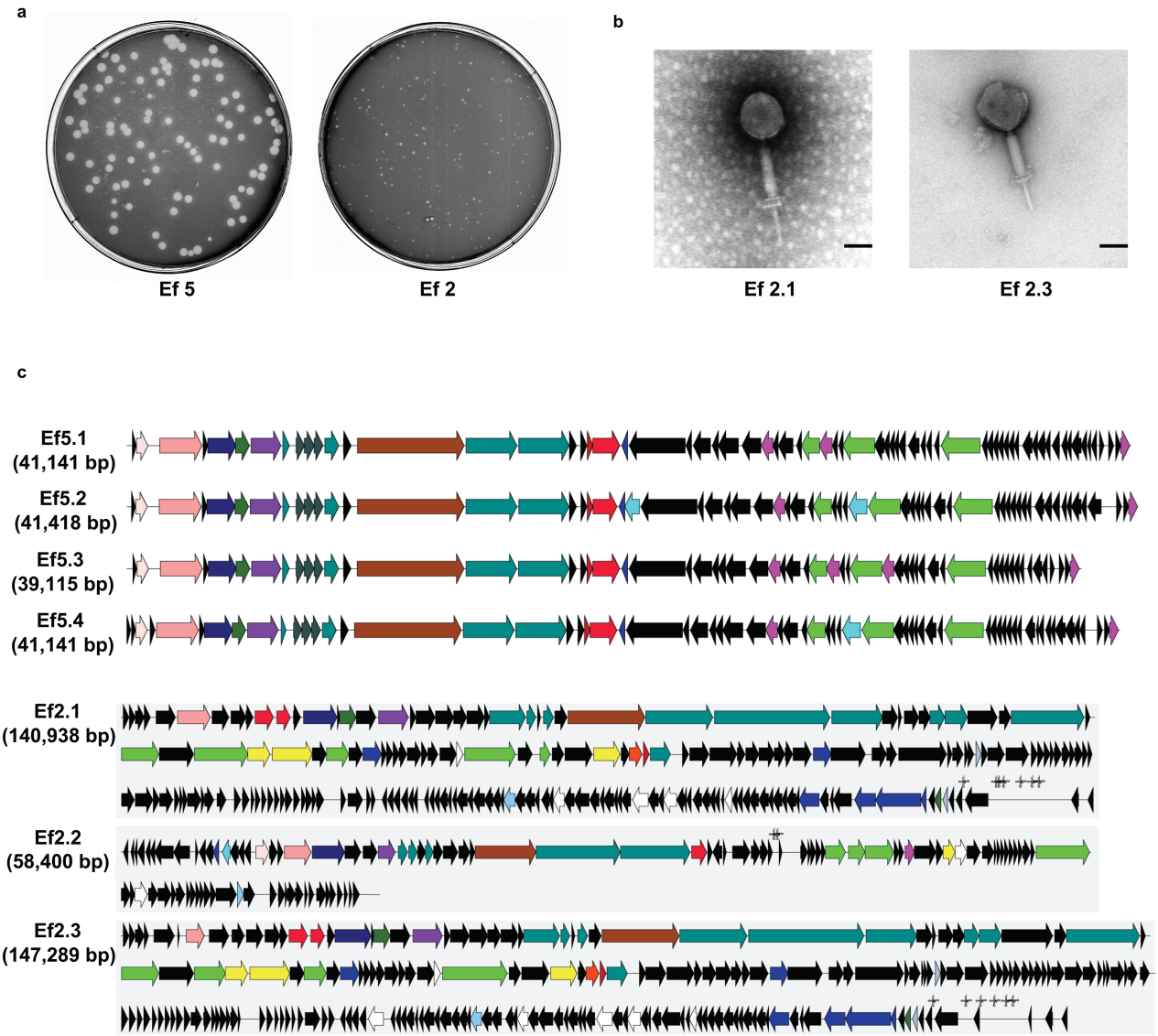
Capsid structural components and assembly	Tail and baseplate assembly	DNA metabolism	Host cell lysis	Other
encapsidation/terminase major capsid portal	tail fiber tail components	DNA polymerase DNA binding	endolysin/lysin holin	hypothetical protein intron/intein

Extended Data Fig. 4 | Isolation and amplification of phages against cytolytic *E. faecalis* isolated from mice. a, BHI agar plate showing phage plaque morphology. The phage cocktail (100 µl) (10^2–10^3 PFUs) was mixed with overnight-grown *E. faecalis* culture (100 µl) and then added to BHI broth top agar (0.5% agar) and poured over a BHI plate (1.5% agar). After overnight growth at 37 °C, images were captured on an Epson Perfection 4990 Photo scanner. **b**, Simplified illustration of the morphologies of different phages. Siphophages have long, flexible noncontractile tails (left); myophages have contractile tails (middle); and podophages have short noncontractile tails (right). **c**, Transmission electron microscopy revealed that phages we isolated were all podophages (Efmus1, Efmus2, Efmus3 and Efmus4). **d**, Genetic map of phage genomes. The linear maps are based on nucleotide sequences of the phage genomes and predicted open reading frames. The name and length (in bp) of each genome are indicated to the left of each phage map. Protein-coding sequences are coloured on the basis of functional role categories. Scale bar, 50 nm. All results were generated from at least three independent replicates.



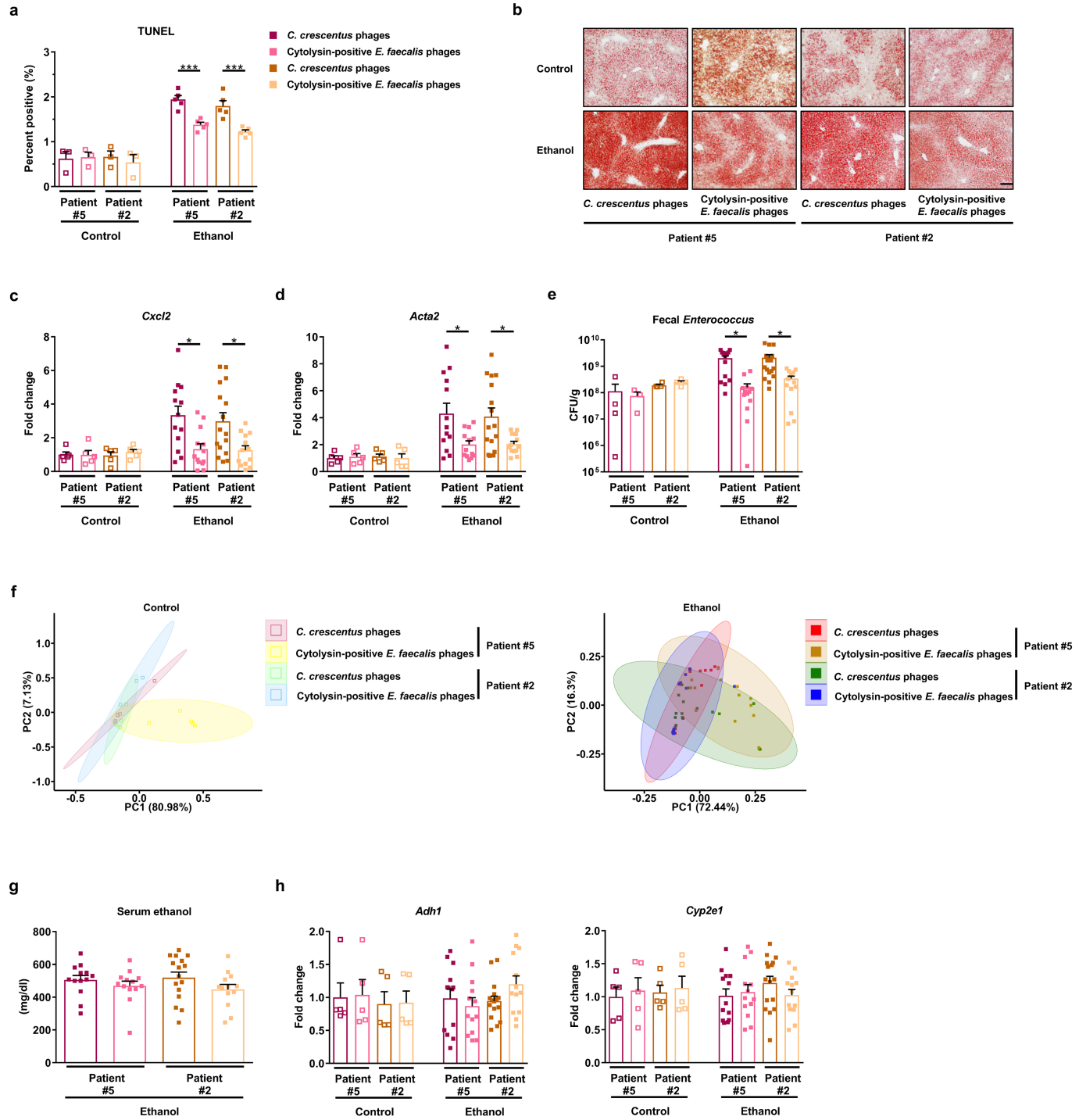
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Phages reduce translocation of cytolysin to the liver and reduce ethanol-induced liver disease in *Atp4a*$^{SI/SI}$ mice. **a**–**k**, Wild-type (WT) and *Atp4a*$^{SI/SI}$ littermates were fed oral isocaloric (control) or chronic–binge ethanol diets, and gavaged with vehicle (PBS), control phages against *C. crescentus* (10^{10} PFUs) or a cocktail of four different phages that target cytolytic *E. faecalis* (10^{10} PFUs), 1 day before an ethanol binge. **a**, Serum levels of ALT. **b**, Hepatic triglyceride content. **c**, Representative oil red O-stained liver sections. **d**–**f**, Hepatic levels of mRNAs. **g**, Proportions of mice that were positive for cytolysin in the liver, measured by qPCR for *cylL*$_S$. **h**, Faecal CFUs of *Enterococcus*. **i**, Faecal samples were collected and 16S rRNA genes were sequenced. PCoA based on Jaccard dissimilarity matrices found no significant difference in faecal microbiota among mice given PBS, control phage or phages that target cytolytic *E. faecalis* in each group. **j**, **k**, Serum levels of ethanol and hepatic levels of *Adh1* and *Cyp2e1* mRNAs did not differ significantly among colonized mice after ethanol feeding. Scale bar, 100 μm. Results are expressed as mean ± s.e.m. (**a**, **b**, **d**–**f**, **h**, **j**, **k**). *P* values were determined by two-way ANOVA with Tukey's post hoc test (**a**, **b**, **d**–**f**, **h**, **j**, **k**), two-sided Fisher's exact test followed by FDR procedures (**g**) or PERMANOVA followed by FDR procedures (**i**). All results were generated from at least three independent replicates. The exact group size (*n*) and *P* values for each comparison are listed in Supplementary Table 10. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$.

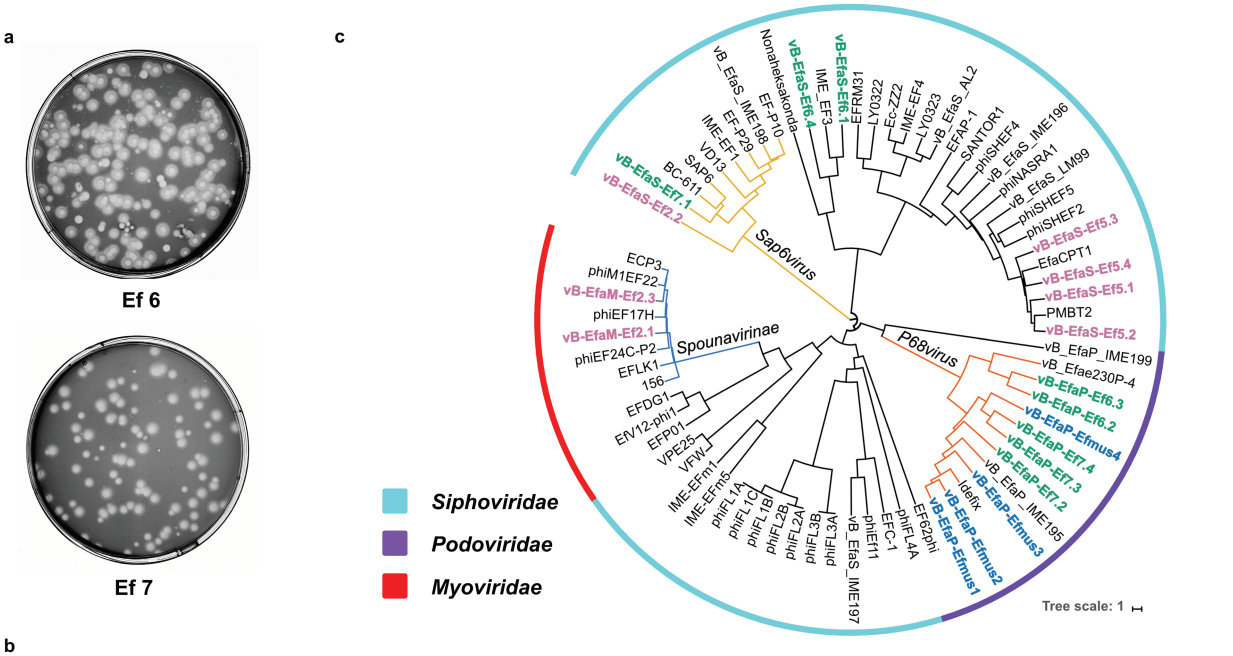


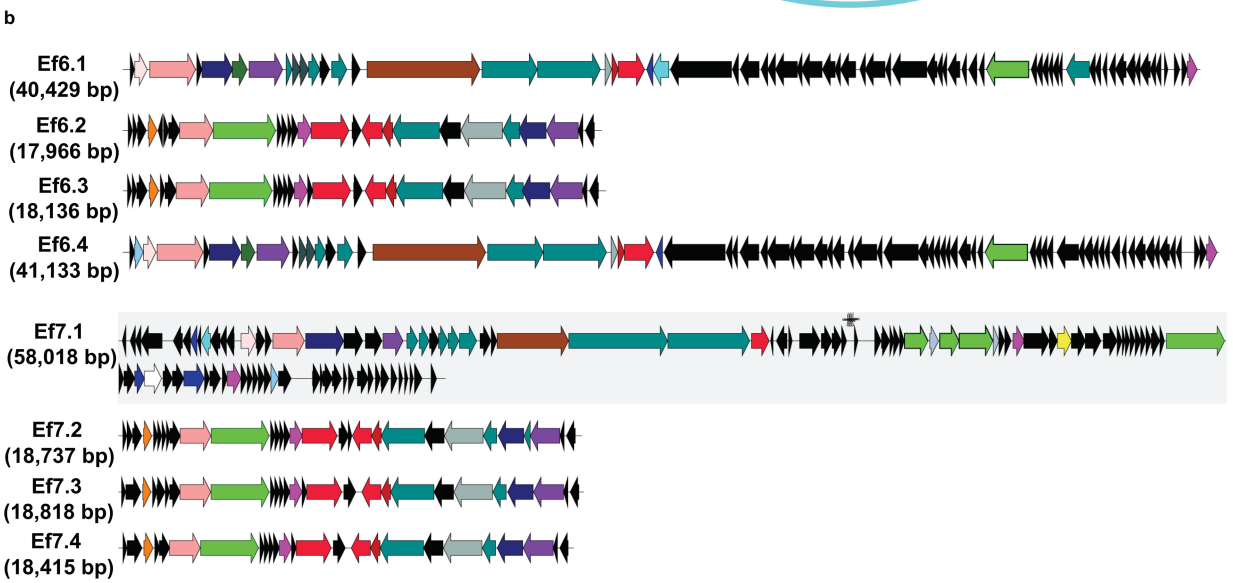
Capsid structural components and assembly	Tail and baseplate assembly	DNA metabolism	Host cell lysis	Other
encapsidation/terminase	tail fibers	DNA polymerase	endolysin/lysin	hypothetical protein
major capsid	tail component	DNA binding	holin	intron/intein
portal	tail measure	DNA recombination		regulatory
protease (scaffold protein)		salvage of nucleosides and nucleotides		transcription factor
other		DNA restriction/ modification		other known functions
		nucleotide degradation		tRNA

Extended Data Fig. 6 | Isolation and amplification of phages against cytolytic *E. faecalis* strains isolated from patients with alcoholic hepatitis. **a**, BHI agar plates showing phage plaque morphology. **b**, Transmission electron microscopy graphs of myophages Ef2.1 and Ef2.3, stained with phosphotungstic acid showing contracted tails. **c**, Genetic map of phage genomes. The linear maps are based on nucleotide sequences of the phage genomes and predicted open reading frames. The name and length (in bp) of each genome are indicated to the left of each phage map. Protein-coding sequences are coloured on the basis of functional role categories. Sequences that encode tRNA genes are indicated by a cloverleaf structure. Scale bar, 50 nm. All results were generated from at least three independent replicates.



Extended Data Fig. 7 | Phages that target cytolytic *E. faecalis* reduce ethanol-induced liver disease in gnotobiotic mice. a–**h**, C57BL/6 germ-free mice were colonized with faeces from two cytolysin-positive patients with alcoholic hepatitis (faeces from one patient were also used in Fig. 2). The mice were then fed oral isocaloric (control) or chronic–binge ethanol diets, and gavaged with control phages against *C. crescentus* (10^{10} PFUs) or a cocktail of 3 or 4 different phages that target cytolytic *E. faecalis* (10^{10} PFUs), one day before an ethanol binge. **a**, Percentage of TUNEL-positive hepatic cells. **b**, Representative oil red O-stained liver sections. **c**, **d**, Hepatic levels of mRNAs that encode the inflammatory cytokine *Cxcl2*, and *Acta2* (a marker of activated hepatic stellate cells). **e**, Faecal CFUs of *Enterococcus*. **f**, Faecal samples were collected and 16S rRNA genes were sequenced. PCoA based on Jaccard dissimilarity matrices shows no significant differences in the faecal microbiota of mice gavaged with control phage and phages that target cytolytic *E. faecalis* in each group. **g**, **h**, Serum levels of ethanol and hepatic levels of *Adh1* and *Cyp2e1* mRNAs did not differ significantly among colonized mice after ethanol feeding. Scale bar, 100 µm. Results are expressed as mean ± s.e.m. (**a**, **c**–**e**, **g**, **h**). *P* values were determined by two-way ANOVA with Tukey's post hoc test (**a**, **c**–**e**, **g**, **h**) or PERMANOVA followed by FDR procedures (**f**). All results were generated from at least three independent replicates. The exact group size (*n*) and *P* values for each comparison are listed in Supplementary Table 10. $^{*}P < 0.05$, $^{***}P < 0.001$.

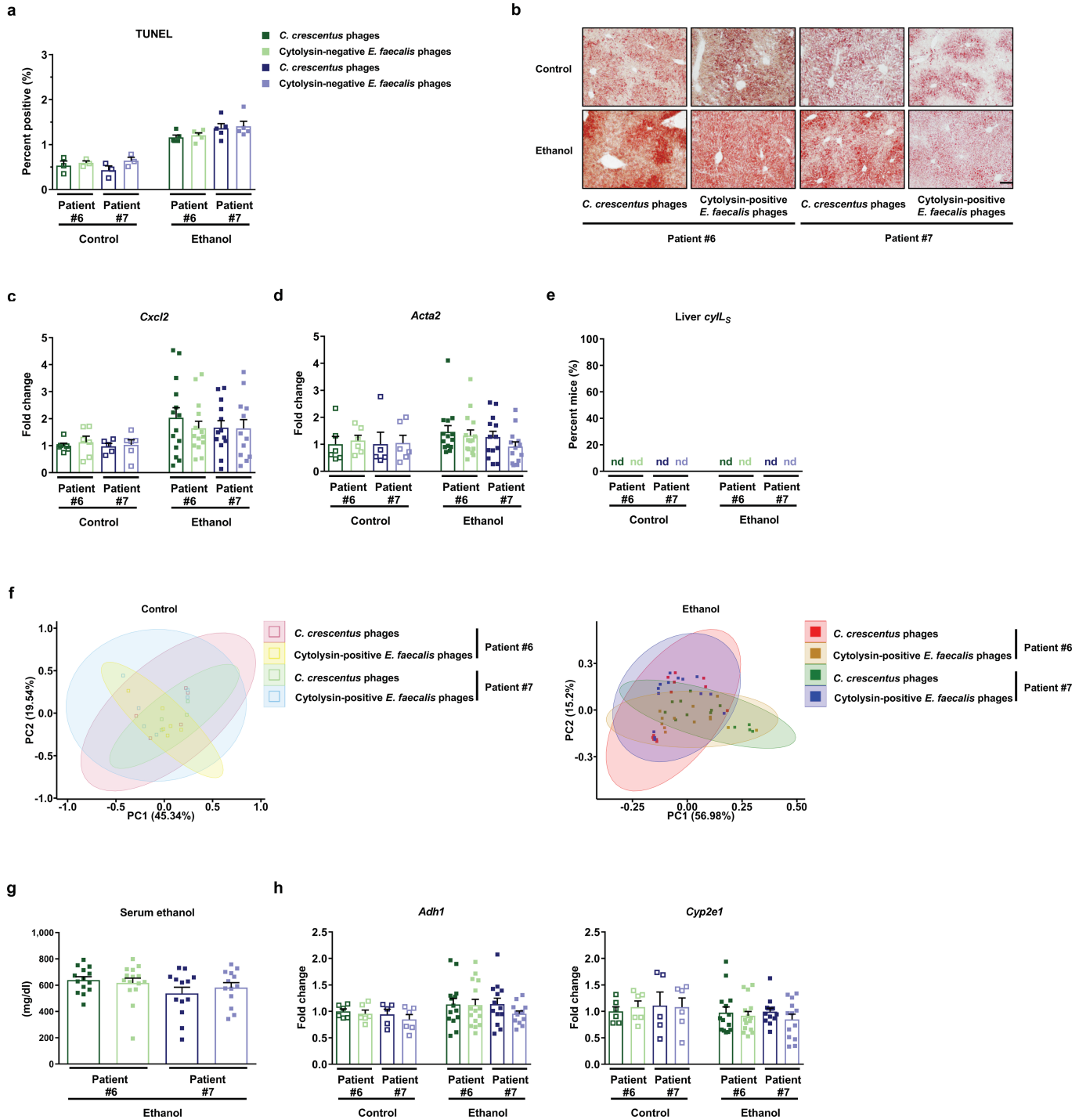




Capsid structural components and assembly	Tail and baseplate assembly	DNA metabolism	Host cell lysis	Other
encapsidation/terminase	tail fibers	DNA polymerase	endolysin/lysin	hypothetical protein
major capsid	tail component	DNA binding	holin	intron/intein
portal	tail measure	DNA recombination		regulatory
protease (scaffold protein)		salvage of nucleosides and nucleotides		tRNA
other		DNA restriction/ modification		
		nucleotide degradation		

Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Isolation and amplification of phages against non-cytolytic *E. faecalis* strains isolated from patients with alcoholic hepatitis. **a**, BHI agar plates showing phage plaque morphology. **b**, Genetic map of phage genomes. The linear maps are based on nucleotide sequences of the phage genomes and predicted open reading frames. The name and length (in bp) of each genome are indicated to the left of each phage map. Protein-coding sequences are coloured on the basis of functional role categories. Sequences that encode tRNA genes are indicated by a cloverleaf structure. **c**, Phylogenetic tree of *Enterococcus* phages. A whole-genome average nucleotide distance tree was constructed for 73 available *Enterococcus* phage genomes: 54 of these were from GenBank (denoted by black letters) and 19 were from this study (4 phages against cytolysin-positive *E. faecalis* isolated from mice (shown in blue letters); 7 phages against cytolysin-positive *E. faecalis* isolated from patients with alcoholic hepatitis (shown in pink letters); and 8 phages against cytolysin-negative *E. faecalis* isolated from patients with alcoholic hepatitis (shown in green letters)) with Mash[43] using a sketch size of $s = 5000$ and a k-mer size of $k = 12$ and GGRaSP[59] (Methods). Coloured branches denote specific phage genera or subfamily: Sap6virus, P68virus and *Spounavirinae*. The scale bar represents per cent average nucleotide divergence. All results were generated from at least three independent replicates.



Extended Data Fig. 9 | Phages that target non-cytolytic *E. faecalis* do not reduce ethanol-induced liver disease in gnotobiotic mice. a–**h**, C57BL/6 germ-free mice were colonized with faeces from two cytolysin-negative patients with alcoholic hepatitis. Transplanted gnotobiotic mice were fed oral isocaloric (control) or chronic–binge ethanol diets and gavaged with control phages against *C. crescentus* (10^{10} PFUs) or a cocktail of four different phages targeting non-cytolytic *E. faecalis* (10^{10} PFUs), 1 day before an ethanol binge. **a**, Percentage of TUNEL-positive hepatic cells. **b**, Representative oil red O-stained liver sections. **c**, **d**, Hepatic levels of mRNAs that encode the inflammatory cytokine *Cxcl2*, and *Acta2* (a marker of activated hepatic stellate cells). **e**, Proportions of mice that were positive for cytolysin in the liver, measured by qPCR for $cylL_S$. **f**, Faecal samples were collected and 16S rRNA genes were sequenced. PCoA based on Jaccard dissimilarity matrices found no significant difference in faecal microbiota among mice gavaged with control phages and phages that target cytolytic *E. faecalis* in each group. **g**, **h**, Serum levels of ethanol and hepatic levels of *Adh1* and *Cyp2e1* mRNAs did not differ significantly among colonized mice after ethanol feeding. Scale bar, 100 µm. Results are expressed as mean ± s.e.m. (**a**, **c**, **d**, **g**, **h**). *P* values were determined by two-way ANOVA with Tukey's post hoc test (**a**, **c**, **d**, **g**, **h**), two-sided Fisher's exact test followed by FDR procedures (**e**) or PERMANOVA followed by FDR procedures (**f**). All results were generated from at least three independent replicates. The exact group size (*n*) and *P* values for each comparison are listed in Supplementary Table 10.

nature research

Corresponding author(s): Bernd Schnabl

Last updated by author(s): Sep 5, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All biochemical assays were measured using SoftMax Pro 7.0.3; qPCRs were run with StepOnePlus real-time PCR system; Liver histological pictures were taken with DP Controller and DP Manager (Olympus); Phage electronic microscopy pictures were taken using Maxim DL5; Plates were scanned using EPSON 4990 Photo; All pictures were viewed using ImageJ
Data analysis	Bacteriophage sequencing and phage tree: Albacore v2.3.4 (ONT), Porechop v0.2.3, Unicycler v0.4.7 pipeline, Pilon v1.22, CLC Genomics Workbench 4.9, NCBI Prokaryotic Genome Annotation Pipeline, in-house PERL script using Xfig, Phage_Finder, MASH program, GGRaSP and APE R-package, iTOL tree viewer 16S sequencing: MOTHUR-based 16S rDNA analysis workflow E. faecalis genome sequencing and tree: abricate v0.8.10 , Prokka, Roary, RAxML, iTOL Statistical analyses: R statistical software 3.5.1, GraphPad Prism v6.01

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw 16S sequence reads can be found in the NCBI SRA associated with Bioproject PRJNA525701. Bacteriophage raw sequence reads and annotated genomes are

available at NCBI under the following consecutive BioSample IDs (SAMN11089809 – SAMN11089827). Genome sequence data of E. faecalis strains isolated in this study were registered at ENA under Study PRJEB25007.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No power analyses or other calculations were used to predetermine sample sizes. Sample sizes were chosen based on prior literature using similar experimental paradigms (Nat Commun. 2017;8:2137; Gut. 2019;68:1504-1515)
Data exclusions	No data were excluded
Replication	In vivo experiments: more than two technical replicates (from different cohorts, on different dates), as well as biological replicates were performed to ensure data reproducibility; In vitro experiments: three independent experiments and also replicates were performed on different dates to ensure data reproducibility. All replications were successful.
Randomization	Mice of similar age and weight were randomly assigned to experimental and control groups.
Blinding	The investigators were not blinded during cell and animal experiment assays.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Female and male C57BL/6 mice (age, 9–12 weeks) (strain: wild type, Atp4asl/sl)
Wild animals	No wild animals were involved in the study.
Field-collected samples	No field-collected samples were involved in the study
Ethics oversight	All animal studies were reviewed and approved by the Institutional Animal Care and Use Committee of the University of California, San Diego.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Alcoholic hepatitis patients were from multiple centers from United States, Mexico and Europe, with the age ranged from 30 to

Population characteristics	75. Alcohol use disorder patients and non-alcoholic controls were from United States and Europe, with the age ranged from 27 to 74. Both genders were included in all populations. Detailed descriptions in Methods, Extended Data Tables 1 and 2
Recruitment	Patients with alcohol use disorder fulfilling the DSM IV criteria (J Abnorm Psychol. 1997;106:545-553) were recruited from an alcohol withdrawal unit in San Diego, USA and Brussels, Belgium where they followed a detoxification and rehabilitation program. Alcoholic hepatitis patients were enrolled from the InTeam Consortium (ClinicalTrials.gov identifier number: NCT02075918) from centers in the USA, Mexico, United Kingdom, France and Spain. Detailed inclusion and exclusion criteria are listed in Methods
Ethics oversight	The protocol was approved by the Ethics Committee of Hôpital Huriez (Lille, France), Universidad Autonoma de Nuevo Leon (Monterrey, México), Hospital Universitari Vall d'Hebron (Barcelona, Spain), King's College London (London, UK), Yale University (New Haven, USA), University of North Carolina at Chapel Hill (Chapel Hill, USA), Weill Cornell Medical College (New York, USA), Columbia University (New York, USA), University of Wisconsin (Madison, USA), VA San Diego Healthcare System (San Diego, USA) and Université Catholique de Louvain (Brussels, Belgium).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Resolving the fibrotic niche of human liver cirrhosis at single-cell level

https://doi.org/10.1038/s41586-019-1631-3

Received: 4 September 2018

Accepted: 4 September 2019

Published online: 9 October 2019

P. Ramachandran[1]*, R. Dobie[1], J. R. Wilson-Kanamori[1], E. F. Dora[1], B. E. P. Henderson[1], N. T. Luu[2,3], J. R. Portman[1], K. P. Matchett[1], M. Brice[1], J. A. Marwick[1,4], R. S. Taylor[1], M. Efremova[5], R. Vento-Tormo[5], N. O. Carragher[4], T. J. Kendall[1,6], J. A. Fallowfield[1], E. M. Harrison[7], D. J. Mole[1,7], S. J. Wigmore[1,7], P. N. Newsome[2,3], C. J. Weston[2,3], J. P. Iredale[8], F. Tacke[9], J. W. Pollard[10,11], C. P. Ponting[12], J. C. Marioni[5,13,14], S. A. Teichmann[5,13,15] & N. C. Henderson[1]*

Liver cirrhosis is a major cause of death worldwide and is characterized by extensive fibrosis. There are currently no effective antifibrotic therapies available. To obtain a better understanding of the cellular and molecular mechanisms involved in disease pathogenesis and enable the discovery of therapeutic targets, here we profile the transcriptomes of more than 100,000 single human cells, yielding molecular definitions for non-parenchymal cell types that are found in healthy and cirrhotic human liver. We identify a scar-associated TREM2$^+$CD9$^+$ subpopulation of macrophages, which expands in liver fibrosis, differentiates from circulating monocytes and is pro-fibrogenic. We also define ACKR1$^+$ and PLVAP$^+$ endothelial cells that expand in cirrhosis, are topographically restricted to the fibrotic niche and enhance the transmigration of leucocytes. Multi-lineage modelling of ligand and receptor interactions between the scar-associated macrophages, endothelial cells and PDGFRα$^+$ collagen-producing mesenchymal cells reveals intra-scar activity of several pro-fibrogenic pathways including TNFRSF12A, PDGFR and NOTCH signalling. Our work dissects unanticipated aspects of the cellular and molecular basis of human organ fibrosis at a single-cell level, and provides a conceptual framework for the discovery of rational therapeutic targets in liver cirrhosis.

Recent estimates suggest that 844 million people worldwide have chronic liver disease, with two million deaths per year and a rising incidence[1]. Iterative liver injury secondary to any cause leads to progressive fibrosis and ultimately results in liver cirrhosis. Notably, the degree of liver fibrosis predicts adverse patient outcomes[2]. Hence, effective antifibrotic therapies for patients with chronic liver disease are urgently required[3,4].

Liver fibrosis involves a complex interplay between multiple non-parenchymal cell (NPC) lineages including immune, endothelial and mesenchymal cells spatially located within areas of scarring, termed the fibrotic niche. Despite progress in our understanding of liver fibrogenesis accrued using rodent models, there remains a considerable 'translational gap' between putative targets and effective patient therapies[3,4]. This is in part due to limited definition of the functional heterogeneity and interactome of cell lineages that contribute to the fibrotic niche of human liver cirrhosis, which is imperfectly recapitulated by rodent models[3].

Single-cell RNA sequencing (scRNA-seq) is delivering a step change in our understanding of disease pathogenesis, allowing the interrogation of individual cell populations at unprecedented resolution[5]. Here, we studied the mechanisms that regulate human liver fibrosis using scRNA-seq.

Single-cell atlas of human liver NPCs

Hepatic NPCs were isolated from healthy and cirrhotic human livers spanning a range of aetiologies of cirrhosis (Fig. 1a, Extended Data Fig. 1a). Leucocytes (CD45$^+$) or other NPC (CD45$^-$) fractions (Extended Data Fig. 1b) were sorted by flow cytometry before scRNA-seq analysis. To discriminate between liver-resident and circulating leucocytes, we also performed scRNA-seq on CD45$^+$CD66b$^-$ peripheral blood mononuclear cells (PBMCs) (Extended Data Fig. 1c, g–i). The combined tissue and PBMC dataset was partitioned into clusters (Extended Data Fig. 1d) and annotated using signatures of known lineage markers (Extended

[1]University of Edinburgh Centre for Inflammation Research, The Queen's Medical Research Institute, Edinburgh BioQuarter, Edinburgh, UK. [2]NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust and University of Birmingham, Birmingham, UK. [3]Institute of Immunology and Immunotherapy, University of Birmingham, Birmingham, UK. [4]Cancer Research UK Edinburgh Centre, MRC Institute of Genetics and Molecular Medicine at the University of Edinburgh, Edinburgh, UK. [5]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [6]Division of Pathology, University of Edinburgh, Edinburgh, UK. [7]Clinical Surgery, University of Edinburgh, Royal Infirmary of Edinburgh, Edinburgh, UK. [8]Office of the Vice Chancellor, Beacon House and National Institute for Health Research, Biomedical Research Centre, Bristol, UK. [9]Department of Hepatology and Gastroenterology, Charité University Medical Center, Berlin, Germany. [10]MRC Centre for Reproductive Health, The Queen's Medical Research Institute, University of Edinburgh, Edinburgh, UK. [11]Department of Developmental and Molecular Biology, Albert Einstein College of Medicine, New York, NY, USA. [12]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine at the University of Edinburgh, Edinburgh, UK. [13]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, UK. [14]Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Cambridge, UK. [15]Theory of Condensed Matter Group, The Cavendish Laboratory, University of Cambridge, Cambridge, UK. *e-mail: prakash.ramachandran@ed.ac.uk; neil.henderson@ed.ac.uk

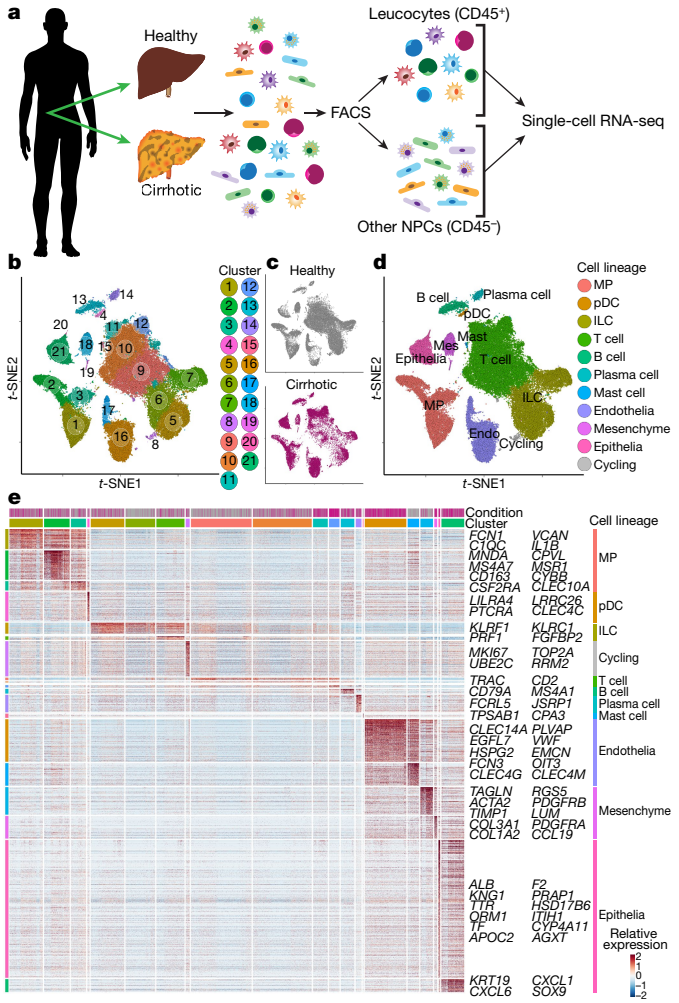


Fig. 1 | Single-cell atlas of human liver NPCs. a, Overview, illustrating the isolation, FACS sorting and scRNA-seq analysis of leucocytes (CD45$^+$) and other NPC fractions (CD45$^-$). **b**, Clustering of 66,135 cells from healthy ($n = 5$) and cirrhotic ($n = 5$) human livers. **c**, Annotation by injury condition. **d**, Cell lineage inferred from expression of marker gene signatures. ILC, innate lymphoid cell; MP, mononuclear phagocyte; pDC, plasmacytoid dendritic cell. **e**, Heat map of cluster marker genes (top, colour-coded by cluster and condition), with cell lineage of exemplar genes labelled (right). Columns denote cells; rows denote genes.

Data Fig. 1d, e, Supplementary Table 1). To generate an atlas of liver-resident cells, contaminating circulating cells were removed from the liver tissue datasets, by excluding cells from the tissue samples which mapped transcriptionally to blood-derived clusters 1 and 13 (Extended Data Fig. 1d). Liver-resident cells expressed higher levels of tissue-residency markers such as CXCR4 compared with PBMCs (Extended Data Fig. 1f).

Re-clustering the 66,135 liver-resident cells from 10 livers ($n = 5$ healthy and $n = 5$ cirrhotic) revealed 21 populations (Fig. 1b), each containing cells from both healthy and cirrhotic livers (Fig. 1c, Extended Data Fig. 2), across 10 cell lineages (Fig. 1d, Extended Data Fig. 2a, b). Subpopulation markers were identified across all clusters and lineages (Fig. 1e, Supplementary Tables 3, 4). Quality control metrics were highly reproducible between individual samples and conditions (Extended Data Fig. 2c–f, Supplementary Table 2). Expression of collagens type I and III, the main fibrillar collagens within the fibrotic niche, was restricted to cells of the mesenchymal lineage (Fig. 1e).

We proceeded to annotate all human liver NPC lineages (below, Supplementary Notes 1–3, Extended Data Fig. 3), and provide an open-access gene browser (http://www.livercellatlas.mvm.ed.ac.uk) that allows assessment of NPC gene expression between healthy and cirrhotic livers.

Distinct macrophages inhabit the fibrotic niche

Previous studies in rodents have highlighted macrophage subpopulations that orchestrate both the progression and regression of liver fibrosis[6–8]. Clustering of mononuclear phagocytes (MPs) identified ten clusters; annotated as scar-associated macrophages (SAMacs), Kupffer cells (KCs), tissue monocytes (TMs), conventional dendritic cells (cDCs) and cycling (proliferating) cells (Fig. 2a, Extended Data Fig. 4a, Supplementary Note 2). Clusters MP(4) and MP(5)—named SAMac(1) and SAMac(2), respectively—were expanded in cirrhotic livers (Fig. 2b), as confirmed by quantification of the MP cell composition of each liver individually (Fig. 2c).

Clusters MP(6) and MP(7) were enriched in the expression of *CD163*, *MARCO* and *TIMD4* (Extended Data Fig. 4b); tissue staining confirmed these as KCs (resident liver macrophages), facilitating the annotation of these clusters as KC(1) and KC(2), respectively (Extended Data Fig. 4c). A lack of *TIMD4* expression distinguished cluster KC(2) from KC(1) (Extended Data Fig. 4b); cell counting demonstrated TIMD4$^+$ cell numbers to be equivalent between healthy and cirrhotic livers, but showed a loss of MARCO$^+$ cells, consistent with a selective reduction in MARCO$^+$TIMD4$^-$ KC(2) in liver fibrosis (Fig. 2c, Extended Data Fig. 4d, e).

Scar-associated clusters SAMac(1) and SAMac(2) expressed the unique markers *TREM2* and *CD9* (Fig. 2d, e). These macrophages displayed a hybrid phenotype, with features of both TMs and KCs (Fig. 2d, e), analogous to monocyte-derived macrophages in mouse liver injury models[7]. Flow cytometry confirmed expansion of TREM2$^+$CD9$^+$ macrophages in human fibrotic livers (Fig. 2f, Extended Data Fig. 4f). Conditioned medium from SAMacs after fluorescence-activated cell sorting (FACS) promoted fibrillar collagen expression by primary human hepatic stellate cells (HSCs) (Fig. 2g), indicating that SAMacs have a pro-fibrogenic phenotype. Tissue staining demonstrated the presence of TREM2$^+$CD9$^+$MNDA$^+$ SAMacs topographically localized in collagen-positive scar regions (Fig. 2h, Extended Data Fig. 4g–i), and significantly expanded in cirrhotic livers (Extended Data Fig. 4j, k). Cell counting of stained cirrhotic livers morphologically segmented into regions of fibrotic septae and parenchymal nodules, confirmed SAMac accumulation within the fibrotic niche (Extended Data Fig. 4l).

Local proliferation has a significant role in macrophage expansion at sites of fibrosis in rodent models[7,9]. Cycling MP cells (Fig. 2a) subclustered into subpopulations of conventional dendritic cells (cDC1 and cDC2), KCs and SAMacs (Extended Data Fig. 4m, Supplementary Table 8). Cycling SAMacs expanded in cirrhosis (Extended Data Fig. 4m), which highlights the potential role of macrophage proliferation in promoting SAMac accumulation in the fibrotic niche.

Pro-fibrogenic phenotype of SAMacs

To delineate the functional profile of SAMacs, we visualized co-ordinately expressed gene groups across the MP subpopulations using self-organizing maps (Extended Data Fig. 5a). We identified six optimally differentiating metagene signatures, denoted as A–F (Extended Data Fig. 5a, Supplementary Table 9). Signatures A and B defined SAMacs and were enriched for ontology terms relevant to tissue fibrosis (Extended Data Fig. 5b). These SAMac-defining signatures included genes such as *TREM2*, *IL1B*, *SPP1*, *LGALS3*, *CCR2* and *TNFSF12*, some of which are known to regulate the function of scar-producing myofibroblasts in fibrotic liver diseases[10–13]. The remaining MP subpopulations were defined by signature C (KCs), signatures D, E (TMs) and signature F (cDC1); ontology terms matched known functions for the associated cell type (Extended Data Fig. 5b, Supplementary Table 9).

In mice, under homeostatic conditions, embryologically derived self-renewing tissue-resident KCs predominate[14–16]. However, after

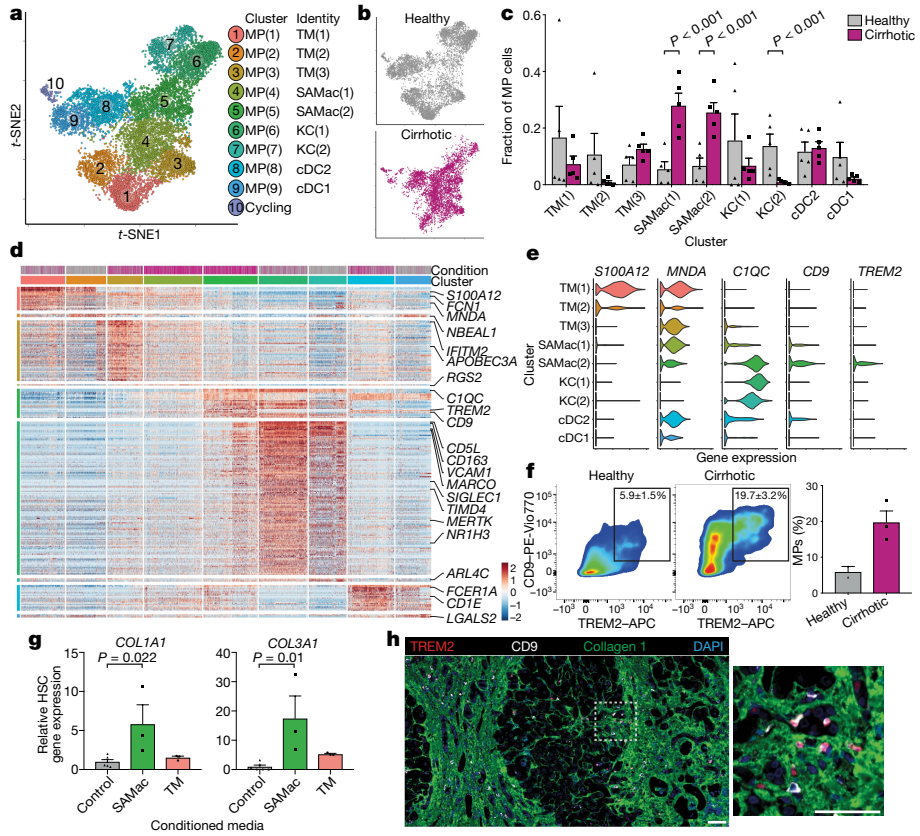


Fig. 2 | Identifying SAMac subpopulations. a, Clustering of 10,737 MPs from healthy ($n = 5$) and cirrhotic ($n = 5$) human livers. **b**, Annotation by injury condition. **c**, Fractions of MP subpopulations in healthy ($n = 5$) and cirrhotic ($n = 5$) livers. **d**, Heat map of MP cluster marker genes (top, colour-coded by cluster and condition), with exemplar genes labelled (right). Columns denote cells; rows denote genes. **e**, Scaled gene expression of SAMac and TM cluster markers across MP cells from healthy ($n = 5$) and cirrhotic ($n = 5$) livers. **f**, Flow cytometry analysis of TREM2$^+$CD9$^+$ MP fraction in healthy ($n = 2$) and cirrhotic ($n = 3$) liver. **g**, Primary human HSCs were treated with conditioned medium from SAMacs ($n = 3$) or TMs ($n = 3$), and indicated genes were analysed by quantitative PCR (qPCR). Expression is shown relative to mean expression of untreated control HSCs ($n = 6$). **h**, Representative immunofluorescence images ($n \geq 3$) of TREM2 (red), CD9 (white), collagen 1 (green) and DAPI (blue) in cirrhotic liver. All scale bars, 50 μm. Data are mean ± s.e.m. *P* values determined by Wald test (**c**) or Kruskal–Wallis and Dunn tests (**g**).

injury, macrophages derived from circulating monocytes accumulate in the liver and regulate fibrosis[7,8]. The ontogeny of human hepatic macrophage subpopulations is unknown. TREM2$^+$CD9$^+$ SAMacs demonstrated a monocyte-like morphology (Fig. 2h, Extended Data Fig. 4g–i) and a distinct topographical distribution from KCs (Extended Data Fig. 4l). To assess the origin of SAMacs, we performed in silico trajectory analysis on a combined dataset of peripheral blood monocytes and liver-resident MPs. We visualized the transcriptional profile of these cells (Fig. 3a, Extended Data Fig. 5c), mapped them along a pseudotemporal trajectory and interrogated their directionality via spliced and unspliced mRNA ratios (RNA velocity[17]). These analyses suggested a differentiation trajectory from peripheral blood monocytes into either SAMacs or cDCs, with no differentiation from KCs to SAMacs, and no progression from SAMacs to KCs (Fig. 3a, Extended Data Fig. 5c). Additional RNA velocity analyses[17] showed downregulation (negative velocity) of the monocyte gene *MNDA* in SAMacs, upregulation (positive velocity) of the SAMac marker gene *CD9* in TMs, and a lack of KC gene *TIMD4* velocity in SAMacs (Extended Data Fig. 5d). Furthermore, assessment of the probabilities of cells in this dataset transitioning into SAMacs indicated a higher likelihood of TMs than KCs differentiating into SAMacs (Fig. 3b). Overall, these data suggest that SAMacs are monocyte-derived, and represent a terminally differentiated cell state within the fibrotic niche.

To characterize the SAMac phenotype further, we identified differentially expressed genes along monocyte differentiation trajectories. We defined three gene co-expression modules, with module 1 representing upregulated genes during blood monocyte-to-SAMac differentiation (Fig. 3c). Module 1 contained multiple pro-fibrogenic genes including *SPP1*, *LGALS3*, *CCL2*, *CXCL8*, *PDGFB* and *VEGFA*[10–13], and displayed ontology terms that are consistent with the promotion of tissue fibrosis and angiogenesis (Fig. 3c, d, Supplementary Table 10). Module 2 contained genes that were downregulated during the differentiation of monocytes to SAMacs (Fig. 3c, Extended Data Fig. 5e), whereas module 3 encompassed a group of upregulated genes during the differentiation from monocytes to cDCs (Fig. 3c, Extended Data Fig. 5f, Supplementary Table 10). SAMacs isolated from cirrhotic human livers (Fig. 2f, Extended Data Fig. 4f) demonstrated enhanced protein secretion of several of the mediators identified by transcriptional analysis (Extended Data Fig. 5g) and promoted fibrillar collagen expression by primary human HSCs (Fig. 2g), which confirms that SAMacs have a pro-fibrogenic phenotype.

To enable cross-species comparison, we performed scRNA-seq on liver MP cells isolated from control mice or mice treated with chronic carbon tetrachloride (CCl_4)—a mouse model of liver fibrosis[7]. MP cells from fibrotic livers were isolated 24 h after the final CCl_4 injection, a time of active fibrogenesis[7]. Five MP cell clusters were defined (Extended Data Fig. 6a–d, Supplementary Table 11), and injury-specific cluster mMP(2) was differentiated by high expression of *Cd9*, *Trem2*, *Spp1* and *Lgals3* (Extended Data Fig. 6a–d). We confirmed expansion of this CD9$^+$ mSAMac population in liver fibrosis (Extended Data Fig. 6e, f) and co-culture of mSAMacs with quiescent primary mouse HSCs promoted fibrillar collagen expression in HSCs (Extended Data Fig. 6g). Canonical

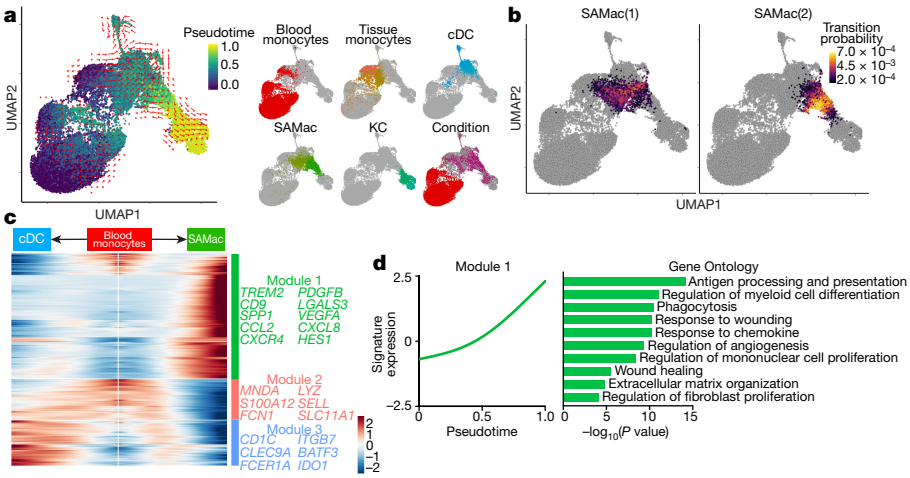


Fig. 3 | Pro-fibrogenic phenotype of SAMacs. a, Uniform manifold approximation and projection (UMAP) visualization of 23,075 cells from liver-resident MPs (healthy, $n = 5$; cirrhotic, $n = 5$) and blood monocytes (PBMCs, $n = 5$), annotating monocle pseudotemporal dynamics (purple to yellow). RNA velocity field (red arrows) visualized using Gaussian smoothing on regular grid. Right, annotation of MP subpopulations and injury condition. **b**, Transition probabilities per SAMac subpopulation, indicating for each cell the likelihood of transition into either SAMac(1) or SAMac(2), calculated using RNA velocity (yellow, high; purple, low; grey, below threshold of 2×10^{-4}). **c**, Heat map with spline curves fitted to genes differentially expressed across blood monocyte-to-SAMac (right arrow) and blood monocyte-to-cDC (left arrow) pseudotemporal trajectories, grouped by hierarchical clustering ($k = 3$). Gene co-expression modules (colour) and exemplar genes from each module are labelled (right). **d**, Spline curve fitted to averaged expression of all genes in module 1 along the monocyte-to-SAMac pseudotemporal trajectory (left), with selected enrichment of Gene Ontology terms (right). *P* values determined by Fisher's exact test.

correlation analysis between human and mouse MP datasets[18] demonstrated that human and mouse SAMacs clustered together (Extended Data Fig. 6h, i) and that this cluster was enriched for SAMac markers *CD9*, *TREM2* and *SPP1* (Extended Data Fig. 6j), confirming that mouse SAMacs represent a corollary population to human SAMacs.

To identify potential transcriptional regulators of human SAMacs, we defined sets of genes co-expressed with known transcription factors (regulons) along the tissue monocyte-to-macrophage pseudotemporal trajectory and in KCs (Extended Data Fig. 5g, h, Supplementary Table 12). This identified regulons and corresponding transcription factors associated with distinct macrophage phenotypes, highlighting *HES1* and *EGR2* activity in SAMacs.

To determine whether SAMacs expand in earlier-stage human liver disease, we analysed cohorts of patients with non-alcoholic fatty liver disease (NAFLD). Application of differential gene expression signatures of human SAMacs, KCs and TMs to a deconvolution algorithm[19] enabled the assessment of hepatic monocyte–macrophage composition in whole liver microarray data across the spectrum of early-stage NAFLD[20] (Extended Data Fig. 7a). This demonstrated expansion of SAMacs in patients with non-alcoholic steatohepatitis (NASH) (Extended Data Fig. 7a, b), an increased frequency of SAMacs with worsening histological NAFLD activity score (NAS) and fibrosis score (Extended Data Fig. 7c), but no association with other patient demographics (Extended Data Fig. 7d). In a separate NAFLD biopsy cohort, the expansion of SAMacs increased with NAFLD activity (Extended Data Fig. 7e) and positively correlated with the degree of fibrosis across the full severity spectrum of NAFLD-induced liver fibrosis (Extended Data Fig. 7f).

In summary, these data demonstrate that TREM2$^+$CD9$^+$ SAMacs derive from the recruitment and differentiation of circulating monocytes, are conserved across species, display a pro-fibrogenic phenotype and expand early in the course of liver disease progression.

Endothelial subpopulations inhabit the fibrotic niche

In rodent models, hepatic endothelial cells are known to regulate fibrogenesis. Clustering of human liver endothelial cells identified seven subpopulations (Fig. 4a). Classical endothelial cell markers did not discriminate between the seven clusters, although Endo(1) was distinct in lacking *CD34* expression (Extended Data Fig. 8a). To annotate endothelial subpopulations fully (Supplementary Note 3, Extended Data Fig. 8k), we identified differentially expressed markers (Fig. 4c, Supplementary Table 13), determined functional expression profiles (Extended Data Fig. 8g, Supplementary Table 14), performed analysis

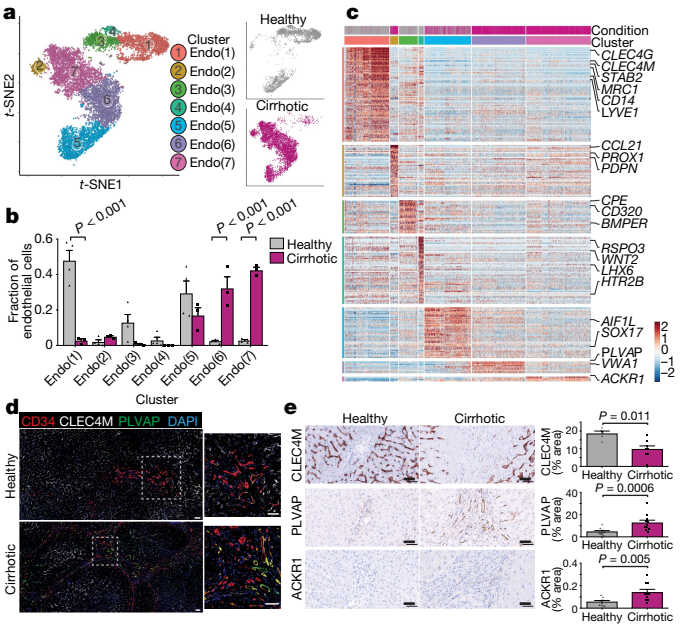


Fig. 4 | Identifying scar-associated endothelial subpopulations. a, Clustering 8,020 endothelial cells from healthy ($n = 4$) and cirrhotic ($n = 3$) human livers, annotating injury condition (right). **b**, Fractions of endothelial subpopulations in healthy ($n = 4$) and cirrhotic ($n = 3$) livers. **c**, Heat map of endothelial cluster marker genes (colour-coded by cluster and condition), with exemplar genes labelled (right). Columns denote cells; rows denote genes. **d**, Representative immunofluorescence images ($n \geq 3$) of CD34 (red), CLEC4M (white), PLVAP (green) and DAPI (blue) in healthy and cirrhotic human liver. **e**, Digital pixel quantification of CLEC4M staining in healthy ($n = 5$) and cirrhotic ($n = 8$) liver, PLVAP staining in healthy ($n = 11$) and cirrhotic ($n = 11$) liver, and ACKR1 staining in healthy ($n = 10$) and cirrhotic ($n = 10$) liver. All scale bars, 50 µm. Data are mean ± s.e.m. *P* values determined by Wald test (**b**) or two-tailed Mann–Whitney test (**e**).

of transcription factor regulons (Extended Data Fig. 8h, Supplementary Table 15) and assessed spatial distribution via tissue staining (Fig. 4d, Extended Data Fig. 8j).

Disease-specific endothelial cells Endo(6) and Endo(7) (CD34$^+$PLVAP$^+$VWA1$^+$ and CD34$^+$PLVAP$^+$ACKR1$^+$, respectively; Fig. 4a–c, Extended Data Fig. 8b) expanded in cirrhotic liver tissue (Fig. 4e) and were restricted to the fibrotic niche (Fig. 4d, e, Extended Data Fig. 8c), allowing annotation as scar-associated endothelia SAEndo(1) and SAEndo(2), respectively. By contrast, CD34$^-$CLEC4M$^+$ Endo(1) (annotated as liver sinusoidal endothelial cells), were reduced in cirrhotic livers (Fig. 4b, e). Metagene signature analysis demonstrated that Endo(6) (SAEndo(1)) cells expressed pro-fibrogenic genes including *PDGFD*, *PDGFB*, *LOX* and *LOXL2*; associated ontology terms included extracellular matrix organization (signature A; Extended Data Fig. 8g). Endo(7) (SAEndo(2)) cells displayed an immunomodulatory phenotype (signature B; Extended Data Fig. 8g). The most discriminatory marker for this cluster, *ACKR1*, has a role in regulating leucocyte recruitment[21]. We confirmed increased expression of PLVAP, CD34 and ACKR1 on endothelial cells isolated from cirrhotic livers (Extended Data Fig. 8d). Flow-based adhesion assays[22] demonstrated that cirrhotic endothelial cells display enhanced leucocyte transmigration (Extended Data Fig. 8e), which was attenuated by ACKR1 knockdown (Extended Data Fig. 8f).

PDGFRA expression defines SAMes cells

Clustering of human liver mesenchymal cells identified four populations (Fig. 5a, b, Extended Data Fig. 9a, Supplementary Table 16). Cluster Mes(1), distinguished by *MYH11* expression (Fig. 5b, Extended Data Fig. 9a), was identified as vascular smooth muscle cells (VSMCs) (Fig. 5c). Mes(4) demonstrated expression of mesothelial markers (Fig. 5b, Extended Data Fig. 9a). Cluster Mes(2) expressed high levels of *RGS5* (Fig. 5b, Extended Data Fig. 9a), and RGS5 staining identified this population as HSCs (Fig. 5c). RGS5$^+$ cells were absent from the fibrotic niche (Fig. 5c). Cluster Mes(3) (distinguished by *PDGFRA* expression) expressed high levels of fibrillar collagens and pro-fibrogenic genes (Fig. 5b, d, Extended Data Fig. 9a). PDGFRα$^+$ cells expanded in cirrhotic livers (Fig. 5a, e, f) and were mapped to the fibrotic niche (Fig. 5f), enabling annotation as scar-associated mesenchymal (SAMes) cells.

To study SAMes cell heterogeneity, further clustering (Extended Data Fig. 9b) identified two populations of SAMes cells (Extended Data Fig. 9c,d, Supplementary Table 17). *OSR1* expression distinguished cluster SAMesB (Extended Data Fig. 9c), and labelled a subpopulation of periportal cells in healthy liver and scar-associated cells in the fibrotic niche (Extended Data Fig. 9e, f). Cluster SAMesB also expressed other known portal fibroblast markers[23] (Extended Data Fig. 9g).

In rodent liver fibrosis models, HSCs differentiate into scar-producing myofibroblasts[24–26]. Pseudotemporal ordering and RNA velocity analyses demonstrated a trajectory from human HSCs to SAMes cells (Extended Data Fig. 9h). Assessment of gene co-expression modules along the HSC-to-SAMes differentiation continuum indicated upregulation of fibrogenic genes including *COL1A1*, *COL1A2*,*COL3A1* and *TIMP1* and downregulation of genes including *RGS5*, *IGFBP5*, *ADAMTS1* and *GEM*, which are known to be downregulated in mouse HSC in response to liver injury[25] (Extended Data Fig. 9i).

The multi-lineage interactome in the fibrotic niche

Having defined the populations of scar-associated macrophages, endothelial and mesenchymal cells, we confirmed the close topographical association of these cells within the fibrotic niche (Extended Data Fig. 10a, b), and used CellPhoneDB[27] to perform an unbiased ligand–receptor interaction analysis between these populations.

Numerous statistically significant paracrine and autocrine interactions were detected between ligands and cognate receptors expressed

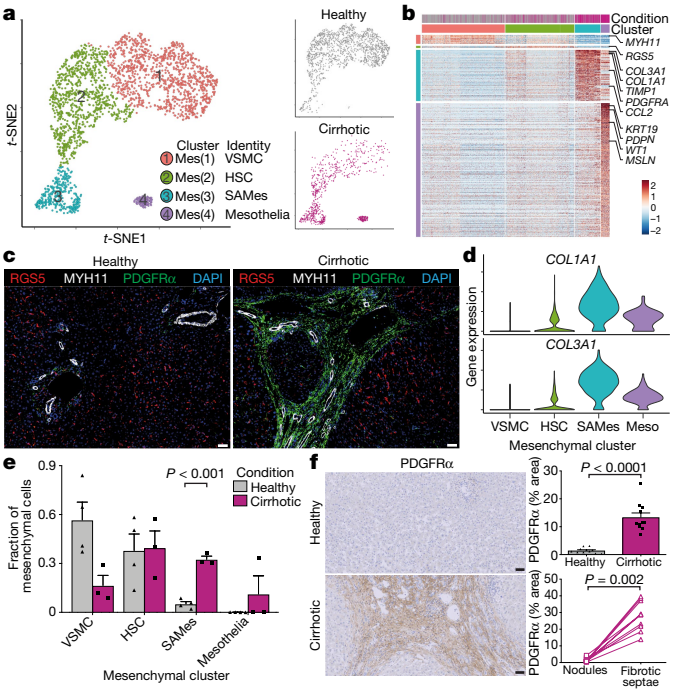


Fig. 5 | Identifying a SAMes subpopulation. a, Clustering of 2,318 mesenchymal cells from healthy ($n = 4$) and cirrhotic ($n = 3$) human livers, annotating injury condition (right). **b**, Heat map of mesenchymal cluster marker genes (top, colour-coded by cluster and condition), with exemplar genes labelled (right). Columns denote cells; rows denote genes. **c**, Representative immunofluorescence images ($n \geq 3$) of RGS5 (red), MYH11 (white), PDGFRα (green) and DAPI (blue) in healthy and cirrhotic human liver. **d**, Scaled gene expression of fibrillar collagens across mesenchymal cells from healthy ($n = 4$) and cirrhotic ($n = 3$) livers. **e**, Fraction of mesenchymal subpopulations in healthy ($n = 4$) and cirrhotic ($n = 3$) livers. **f**, PDGFRα immunohistochemistry (left) and digital pixel quantification (right) in healthy ($n = 11$) and cirrhotic ($n = 11$) livers (top) and in fibrotic septae and parenchymal nodules in cirrhotic livers ($n = 11$; bottom). All scale bars, 50 µm. Data are mean ± s.e.m. *P* values determined by Wald test (**e**), two-tailed Mann–Whitney test (**f**, top), or two-tailed Wilcoxon test (**f**, bottom).

by SAMac, SAEndo and SAMes cells (Supplementary Table 18, Extended Data Fig. 10f–m). To interrogate how scar-associated NPCs regulate fibrosis and to identify tractable therapeutic targets, we focused functional analyses on interactions with SAMes (Fig. 6a, e, Extended Data Fig. 10d). In keeping with our data demonstrating that SAMacs promote fibrillar collagen expression in HSCs (Fig. 2g), SAMacs expressed epidermal growth factor receptor (EGFR) ligands that are known to regulate mesenchymal cell activation[28] (Fig. 6a). In addition, SAMacs expressed the mesenchymal cell mitogens *TNFSF12* and *PDGFB*, signalling to cognate receptors *TNFRSF12A* and *PDGFRA* on SAMes (Fig. 6a). We confirmed localization of these ligand–receptor pairs within the fibrotic niche (Fig. 6b). Both TNFSF12 and PDGF-BB induced proliferation of primary human HSCs, which was inhibited by blockade of TNFRSF12A and PDGFRA, respectively (Fig. 6c, d). Conditioned medium from primary human SAMacs promoted primary human HSC proliferation ex vivo (Extended Data Fig. 10c), demonstrating a functional role for SAMacs in regulating SAMes cell expansion.

SAEndo cells expressed high levels of Notch ligands *JAG1*, *JAG2* and *DLL4* interacting with Notch receptor *NOTCH3* on SAMes cells (Fig. 6e). NOTCH3 was identified on PDGFRα$^+$ SAMes cells within the fibrotic niche (Fig. 6f), and primary endothelial cells from cirrhotic human liver demonstrated increased expression of JAG1 (Fig. 6g). Co-culture of primary human HSCs and endothelial cells from cirrhotic livers promoted fibrillar collagen production by HSCs, which was inhibited by addition

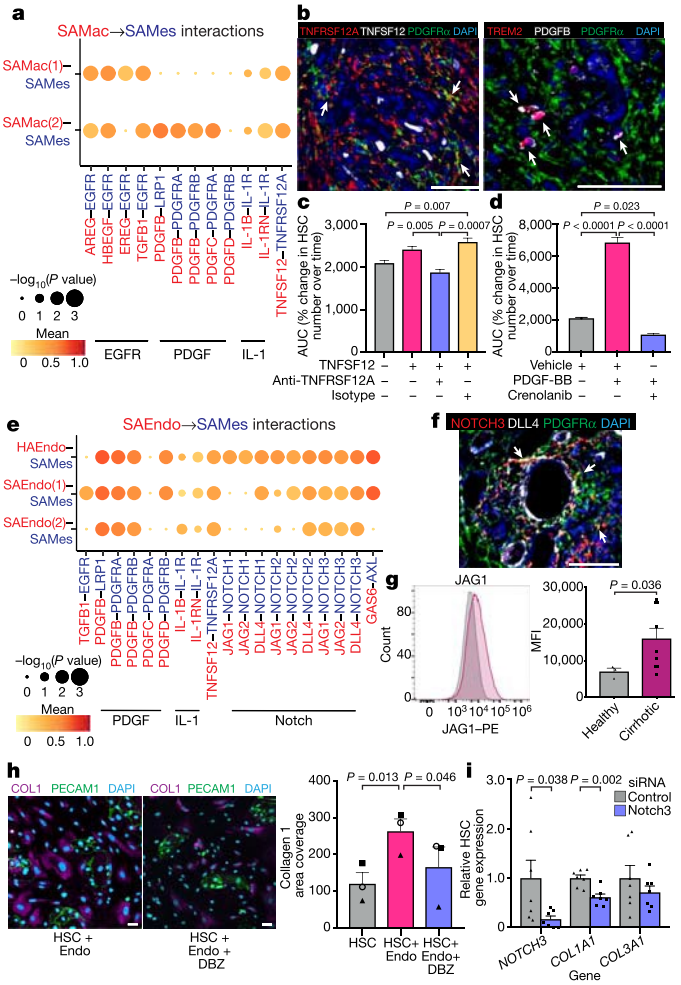


Fig. 6 | Multi-lineage interactions in the fibrotic niche. a, Dot plot of ligand–receptor interactions between SAMac ($n = 10$ human livers) and SAMes ($n = 7$ human livers) subpopulations. Ligand (red) and cognate receptor (blue) are shown on the *x* axis; cell populations that express ligand (red) and receptor (blue) are shown on the *y* axis. Circle size denotes *P* value (permutation test); colour (red, high; yellow, low) denotes average ligand and receptor expression levels in interacting subpopulations. **b**, Left, representative immunofluorescence image ($n \geq 3$) of TNFRSF12A (red), TNFSF12 (white), PDGFRα (green) and DAPI (blue); arrows denote TNFRSF12A$^+$PDGFRα$^+$ cells. Right, representative immunofluorescence image ($n \geq 3$) of TREM2 (red), PDGFB (white), PDGFRα (green) and DAPI (blue); arrows denote TREM2$^+$PDGFB$^+$ cells. Scale bars, 50 µm. **c**, **d**, HSC proliferation assay. The area under the curve (AUC) of the percentage change in HSC number over time (hours) is shown on the *y* axis. $n = 3$ (all conditions in **c** and **d**). **e**, Dot plot of ligand–receptor interactions between SAEndo ($n = 7$ human livers) and SAMes ($n = 7$ human livers) subpopulations as in **a**. **f**, Representative immunofluorescence image ($n \geq 3$) of NOTCH3 (red), DLL4 (white), PDGFRα (green) and DAPI (blue) in the fibrotic niche; arrows denote NOTCH3$^+$PDGFRα$^+$ cells. Scale bar, 50 µm. **g**, Flow cytometry analysis of JAG1 in endothelial cells from healthy ($n = 3$) or cirrhotic ($n = 9$) liver. Left, representative histogram; right, mean fluorescence intensity (MFI). **h**, Co-culture of primary human HSCs and endothelial cells from cirrhotic livers, with or without the Notch inhibitor dibenzazepine (DBZ). Left, representative immunofluorescence images ($n = 3$) of collagen 1 (COL1; magenta), PECAM1 (green) and DAPI (blue). Scale bars, 50 µm. Right, digital pixel analysis of the collagen 1 area ($n = 3$). **i**, Gene knockdown in HSCs using control ($n = 7$) or *NOTCH3* ($n = 7$) siRNA. Indicated genes were analysed by qPCR, with expression relative to mean expression of control siRNA-treated HSCs. Data are mean ± s.e.m. *P* values determined by one-way ANOVA and Tukey test (**c**, **d**), two-tailed Mann–Whitney test (**g**, **i**) or repeated-measures one-way ANOVA and Tukey test (**h**).

of the Notch-signalling inhibitor dibenzazepine (Fig. 6h). Furthermore, knockdown of *NOTCH3* expression in primary human HSCs resulted in reduced fibrillar collagen expression (Fig. 6i).

In summary, our unbiased dissection of the key ligand–receptor interactions between scar-associated NPCs highlights TNFRSF12A, PDGFRA and Notch signalling as important regulators of mesenchymal cell function within the human liver fibrotic niche.

Discussion

Here, using scRNA-seq and spatial mapping, we resolve the fibrotic niche of human liver cirrhosis, identifying pathogenic subpopulations of TREM2$^+$CD9$^+$ macrophages, ACKR1$^+$ and PLVAP$^+$ endothelial cells and PDGFRα$^+$ collagen-producing myofibroblasts. We dissect a complex, pro-fibrotic interactome between multiple scar-associated cell lineages and identify highly relevant intra-scar pathways that are potentially druggable. In this era of precision medicine, this unbiased multi-lineage approach should inform the design of highly targeted combination therapies that will very likely be necessary to achieve effective antifibrotic potency[3,4].

Application of our novel scar-associated cell markers could also potentially inform molecular pathology-based patient stratification, which is fundamental to the prosecution of successful antifibrotic clinical trials. Our work illustrates the power of single-cell transcriptomics to decode the cellular and molecular basis of human organ fibrosis, providing a conceptual framework for the discovery of relevant therapeutic targets to treat patients with a broad range of fibrotic diseases.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at [https://doi.org/10.1038/s41586-019-1631-3].

1. Marcellin, P. & Kutala, B. K. Liver diseases: a major, neglected global public health problem requiring urgent actions and large-scale screening. *Liver Int.* **38** (Suppl. 1), 2–6 (2018).
2. Angulo, P. et al. Liver fibrosis, but no other histologic features, is associated with long-term outcomes of patients with nonalcoholic fatty liver disease. *Gastroenterology* **149**, 389–97.e10 (2015).
3. Ramachandran, P. & Henderson, N. C. Antifibrotics in chronic liver disease: tractable targets and translational challenges. *Lancet Gastroenterol. Hepatol.* **1**, 328–340 (2016).
4. Friedman, S. L., Neuschwander-Tetri, B. A., Rinella, M. & Sanyal, A. J. Mechanisms of NAFLD development and therapeutic strategies. *Nat. Med.* **24**, 908–922 (2018).
5. Stubbington, M. J. T., Rozenblatt-Rosen, O., Regev, A. & Teichmann, S. A. Single-cell transcriptomics to explore the immune system in health and disease. *Science* **358**, 58–63 (2017).
6. Duffield, J. S. et al. Selective depletion of macrophages reveals distinct, opposing roles during liver injury and repair. *J. Clin. Invest.* **115**, 56–65 (2005).
7. Ramachandran, P. et al. Differential Ly-6C expression identifies the recruited macrophage phenotype, which orchestrates the regression of murine liver fibrosis. *Proc. Natl Acad. Sci. USA* **109**, E3186–E3195 (2012).
8. Karlmark, K. R. et al. Hepatic recruitment of the inflammatory Gr1$^+$ monocyte subset upon liver injury promotes hepatic fibrosis. *Hepatology* **50**, 261–274 (2009).
9. Minutti, C. M. et al. Local amplifiers of IL-4Rα-mediated macrophage activation promote repair in lung and liver. *Science* **356**, 1076–1080 (2017).
10. Pradere, J.-P. et al. Hepatic macrophages but not dendritic cells contribute to liver fibrosis by promoting the survival of activated hepatic stellate cells in mice. *Hepatology* **58**, 1461–1473 (2013).
11. Henderson, N. C. et al. Galectin-3 regulates myofibroblast activation and hepatic fibrosis. *Proc. Natl Acad. Sci. USA* **103**, 5060–5065 (2006).
12. Seki, E. et al. CCR2 promotes hepatic fibrosis in mice. *Hepatology* **50**, 185–197 (2009).
13. Syn, W. K. et al. Osteopontin is induced by hedgehog pathway activation and promotes fibrosis progression in nonalcoholic steatohepatitis. *Hepatology* **53**, 106–115 (2011).
14. Scott, C. L. et al. Bone marrow-derived monocytes give rise to self-renewing and fully differentiated Kupffer cells. *Nat. Commun.* **7**, 10321 (2016).
15. Gomez Perdiguero, E. et al. Tissue-resident macrophages originate from yolk-sac-derived erythro-myeloid progenitors. *Nature* **518**, 547–551 (2015).
16. Mass, E. et al. Specification of tissue-resident macrophages during organogenesis. *Science* **353**, aaf4238 (2016).
17. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

18. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
19. Schelker, M. et al. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat. Commun.* **8**, 2032 (2017).
20. Ahrens, M. et al. DNA methylation analysis in nonalcoholic fatty liver disease suggests distinct disease-specific and remodeling signatures after bariatric surgery. *Cell Metab.* **18**, 296–302 (2013).
21. Pruenster, M. et al. The Duffy antigen receptor for chemokines transports chemokines and supports their promigratory activity. *Nat. Immunol.* **10**, 101–108 (2009).
22. Shetty, S., Weston, C. J., Adams, D. H. & Lalor, P. F. A flow adhesion assay to study leucocyte recruitment to human hepatic sinusoidal endothelium under conditions of shear stress. *J. Vis. Exp.* **85**, 51330 (2014).
23. Iwaisako, K. et al. Origin of myofibroblasts in the fibrotic liver in mice. *Proc. Natl Acad. Sci. USA* **111**, E3297–E3305 (2014).
24. Henderson, N. C. et al. Targeting of α_v integrin identifies a core molecular pathway that regulates fibrosis in several organs. *Nat. Med.* **19**, 1617–1624 (2013).
25. De Minicis, S. et al. Gene expression profiles during hepatic stellate cell activation in culture and *in vivo*. *Gastroenterology* **132**, 1937–1946 (2007).
26. Mederacke, I., Dapito, D. H., Affò, S., Uchinami, H. & Schwabe, R. F. High-yield and high-purity isolation of hepatic stellate cells from normal and fibrotic mouse livers. *Nat. Protocols* **10**, 305–315 (2015).
27. Vento-Tormo, R. et al. Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563**, 347–353 (2018).
28. Minutti, C. M. et al. A macrophage-pericyte axis directs tissue restoration via amphiregulin-induced transforming growth factor beta activation. *Immunity* **50**, 645–654.e6 (2019).

METHODS

Study subjects

Local approval for procuring human liver tissue and blood samples for scRNA-seq, flow cytometry and histological analysis was obtained from the NRS BioResource and Tissue Governance Unit (study number SR574), following review at the East of Scotland Research Ethics Service (reference 15/ES/0094). All subjects provided written informed consent. Healthy background non-lesional liver tissue was obtained intraoperatively from patients undergoing surgical liver resection for solitary colorectal metastasis at the Hepatobiliary and Pancreatic Unit, Department of Clinical Surgery, Royal Infirmary of Edinburgh. Patients with a known history of chronic liver disease, abnormal liver function tests or those who had received systemic chemotherapy within the last four months were excluded from this cohort. Cirrhotic liver tissue was obtained intraoperatively from patients undergoing orthotopic liver transplantation at the Scottish Liver Transplant Unit, Royal Infirmary of Edinburgh. Blood from patients with a confirmed diagnosis of liver cirrhosis were obtained from patients attending the Scottish Liver Transplant Unit, Royal Infirmary of Edinburgh. Patients with liver cirrhosis due to viral hepatitis were excluded from the study. Patient demographics are summarized in Extended Data Fig. 1a. Isolation of primary hepatic macrophage subpopulations and endothelial cells from healthy and cirrhotic livers for cell culture and analysis of secreted mediators was performed at the University of Birmingham, UK. Local ethical approval was obtained (reference 06/Q2708/11, 06/Q2702/61) and all patients provided written, informed consent. Liver tissue was acquired from explanted diseased livers from patients undergoing orthotopic liver transplantation, resected liver specimens or donor livers rejected for transplant at the Queen Elizabeth Hospital, Birmingham. For histological assessment of NAFLD biopsies, anonymized unstained formalin-fixed paraffin-embedded liver biopsy sections encompassing the complete NAFLD spectrum were provided by the Lothian NRS Human Annotated Bioresource under authority from the East of Scotland Research Ethics Service REC 1, reference 15/ES/0094.

Human tissue processing

Importantly, to minimize artefacts[29], we developed a rapid tissue processing pipeline, obtaining fresh non-ischaemic liver tissue taken by wedge biopsy before the interruption of the hepatic vascular inflow during liver surgery or transplantation, and immediately processing this for FACS. This enabled a workflow time of under three hours from patient to single-cell droplet encapsulation.

For human liver scRNA-seq and flow cytometry analyses, a wedge biopsy of non-ischaemic fresh liver tissue (2–3 g) was obtained by the operating surgeon. This was immediately placed in HBSS (Gibco) on ice. The tissue was then transported directly to the laboratory and dissociation routinely commenced within 20 min of the liver biopsy. To enable paired histological assessment, a segment of each liver specimen was also fixed in 4% neutral-buffered formalin for 24 h followed by paraffin-embedding. Additional liver samples, obtained via the same method, were fixed in an identical manner and used for further histological analysis. For human macrophage cell sorting and endothelial cell isolation, liver tissue (40 g) was used from cirrhotic patients undergoing orthotopic liver transplantation or control samples from donor liver or liver resection specimens.

Mice

Adult male C57BL/6JCrl mice aged 8–10 weeks were purchased from Charles River. Mice were housed under specific pathogen-free conditions at the University of Edinburgh. All experimental protocols were approved by the University of Edinburgh Animal Welfare and Ethics Board in accordance with UK Home Office regulations. Liver fibrosis was induced with 4 weeks (nine injections) of twice-weekly intraperitoneal CCl_4 at a dose of 0.4 µl g^{-1} body weight, diluted 1:3 in olive oil as previously described[7]. Mice were randomly assigned to receive CCl_4 or to serve as healthy controls. No sample size calculation or blinding was performed. Liver tissue was obtained 24 h after the final CCl_4 injection, a time of active fibrogenesis[7]. Comparison was made to age-matched uninjured mice.

Preparation of single-cell suspensions

For human liver scRNA-seq, liver tissue was minced with scissors and digested in 5 mg ml^{-1} pronase (Sigma-Aldrich, P5147-5G), 2.93 mg ml^{-1} collagenase B (Roche, 11088815001) and 0.019 mg ml^{-1} DNase (Roche, 10104159001) at 37 °C for 30 min with agitation (200–250 r.p.m.), then strained through a 120-µm nybolt mesh along with PEB buffer (PBS, 0.1% BSA and 2mM EDTA) including DNase (0.019 mg ml^{-1}). Thereafter, all processing was done at 4 °C. The cell suspension was centrifuged at 400*g* for 7 min, supernatant removed, cell pellet resuspended in PEB buffer and DNase added (0.019 mg ml^{-1}), followed by additional centrifugation (400*g*, 7 min). Red blood cell lysis was performed (BioLegend, 420301), followed by centrifugation (400*g*, 7 min), resuspension in PEB buffer and straining through a 35-µm filter. Following another centrifugation at 400*g* for 7 min, cells were blocked in 10% human serum (Sigma-Aldrich, H4522) for 10 min at 4 °C before antibody staining.

For human liver macrophage flow cytometry analysis and cell sorting, and for mouse liver macrophage flow cytometry, cell sorting and scRNA-seq, single-cell suspensions were prepared as previously described, with minor modifications[30]. In brief, liver tissue was minced and digested in an enzyme cocktail 0.625 mg ml^{-1} collagenase D (Roche, 11088882001), 0.85 mg ml^{-1} collagenase V (Sigma-Aldrich, C9263-1G), 1 mg ml^{-1} dispase (Gibco, Invitrogen, 17105-041) and 30 U ml^{-1} DNase (Roche, 10104159001) in RPMI-1640 at 37 °C for 20 min (mouse) or 45 min (human) with agitation (200–250 r.p.m.), before being passed through a 100-µm filter. After lysis of red blood cells (BioLegend, 420301), cells were washed in PEB buffer and passed through a 35-µm filter. Before the addition of antibodies, cells from human samples were blocked in 10% human serum (Sigma-Aldrich, H4522) and mouse samples were blocked in anti-mouse CD16/32 antibody (1:100; BioLegend, 101302) and 10% normal mouse serum (Sigma, M5905) for 10 min at 4 °C.

For human PBMC scRNA-seq, 4.9-ml peripheral venous blood samples were collected in EDTA-coated tubes (Sarstedt, S-Monovette 4.9ml K3E) and placed on ice. Blood samples were transferred into a 50-ml Falcon tube. After lysis of red blood cells (BioLegend, 420301), blood samples were then centrifuged at 500*g* for 5 min and supernatant was removed. Pelleted samples were then resuspended in staining buffer (PBS plus 2% BSA; Sigma-Aldrich) and centrifugation was repeated. Samples were then blocked in 10% human serum (Sigma-Aldrich, H4522) in staining buffer on ice for 30 min. Cells were then resuspended in staining buffer and passed through a 35-µm filter before antibody staining.

Flow cytometry and cell sorting

Incubation with primary antibodies was performed for 20 min at 4 °C. All antibodies, conjugates, lot numbers and dilutions used in this study are presented in Supplementary Table 19. After antibody staining, cells were washed with PEB buffer. For human macrophage flow cytometry analysis and cell sorting, cells were then incubated with streptavidin-BV711 for 20 min at 4 °C (BioLegend 405241; 1:200). For human and mouse cell sorting (FACS) and mouse flow cytometry analysis, cell viability staining (DAPI; 1:1,000) was then performed, immediately before acquiring the samples.

Human cell sorting for scRNA-seq was performed on a BD Influx (Becton Dickinson). Viable single $CD45^+$ (leucocytes) or $CD45^-$ (other non-parenchymal cells) cells were sorted from human liver tissue (Extended Data Fig. 1b) and viable $CD45^+$ $CD66b^-$ (PBMC) cells were sorted from peripheral blood (Extended Data Fig. 1c) and processed for droplet-based scRNA-seq.

To generate conditioned medium from cirrhotic liver macrophage subpopulations, cells were sorted on a BD FACSAria Fusion (Becton Dickinson). Sorted SAMacs (viable $CD45^+Lin^-HLA\text{-}DR^+CD14^+CD16^+$ $CD163^-TREM2^+CD9^+$), TMs (viable $CD45^+Lin^-$ $HLA\text{-}DR^+CD14^+CD16^+$

$CD163^-TREM2^-CD9^-$) and KCs (viable $CD45^+Lin^-HLA\text{-}DR^+CD14^+CD16^+CD163^+CD9^-$) were plated in 12-well plates (Corning, 3513) in DMEM (Gibco, 41965039) containing 2% fetal bovine serum (FBS; Gibco, 10500056) at 1×10^6 cells per ml for 24 h at 37°C 5% CO_2. Control wells contained medium alone. Conditioned medium was collected, centrifuged at 400*g* for 10 min, and supernatant was stored at −80 °C.

For human macrophage flow cytometry analysis, after surface antibody staining, cells were stained with Zombie NIR fixable viability dye (BioLegend, 423105) according to the manufacturer's instructions. Cells were washed in PEB then fixed in Intracellular (IC) Fixation Buffer (Thermo Fisher, 00-8222-49) for 20 min at 4 °C. Fixed samples were stored in PEB at 4 °C until acquisition. Flow cytometry acquisition was performed on a six-laser Fortessa flow cytometer (Becton Dickinson). The gating strategy is shown in Extended Data Fig. 4f and Fig. 2f.

Mouse macrophage cell sorting for scRNA-seq and co-culture experiments was performed on a BD FACSAriaII (Becton Dickinson). For scRNA-seq, viable $CD45^+Lin(CD3, NK1.1, Ly6G, CD19)^-$ cells were sorted from healthy ($n = 3$) and CCl_4-treated ($n = 3$) mice and processed for droplet-based scRNA-seq. For transwell co-culture, viable $CD45^+Lin^-CD11b^+F4/80^+TIMD4^-CD9^+$ (SAMacs) or $CD9^-$ (TMs) cells were sorted from CCl_4-treated mice (Extended Data Fig. 6e). Flow cytometry analysis on macrophages from healthy and CCl_4-treated mice was also performed on a BD FACSAriaII (Becton Dickinson), using the same gating strategy (Extended Data Fig. 6e). All flow cytometry data were analysed using FlowJo software (Treestar).

Luminex assay

Detection of CCL2, galectin-3, IL-1β, CXCL8 and osteopontin (SPP1) and CD163 proteins in conditioned medium from human liver macrophage subpopulations was performed using a custom human luminex assay (R&D Systems), according to the manufacturer's protocol. Data were acquired using a Bio-Plex 200 (Bio-Rad) and are presented as MFI for each analyte.

Cell culture

Primary human HSCs (ScienCell, 5300) were cultured in stellate cell medium (SteCM, ScienCell, 5301) on poly-L-lysine (Sigma, P4832)-coated T75 tissue culture flasks, according to the supplier's protocol. All experiments were performed using cells between passages 3 and 5. For assessment of fibrillar collagen gene expression, HSCs were plated at 75,000 cells per well in 24-well plates (Costar, 3524) in HSC medium consisting of DMEM (Gibco, 21969-035) with 20 μM HEPES (Sigma, H3375), 2 mM L-glutamine (Gibco, 25030-024), 1% penicillin streptomycin (Gibco, 15140-122) and 2% FBS (Gibco, 10270). HSCs were serum-starved overnight (in HSC medium without FBS), washed with PBS, then 250 μl of conditioned medium from primary human macrophage subpopulations was added for 24 h. HSCs were obtained for RNA.

Human HSC proliferation

For proliferation assays, after serum starvation, HSCs were obtained using TrypLE Express (Gibco, 12604013), re-suspended in HSC medium at 2.5×10^4 per ml with Incucyte NucLight Rapid Red (Essen Biosciences, 4717) at a dilution of 1:500 and seeded into 384-well plates (Greiner Bio-One, 781090) at 25 μl per well. HSCs were then treated with (1) control medium; (2) PDGF-BB (10 ng ml^{-1}; Peprotech, 100-14B) or TNFSF12 (100 ng ml^{-1}; Peprotech, 310-06-5) with or without the PDGFRα inhibitor crenolanib[31] (1 μM; Cayman Chemicals, CAY1873), anti-TNFRSF12A (2 μg ml^{-1}; Life Technologies, 16-9018-82, clone ITEM-4), mouse IgG2b kappa isotype control antibody (2 μg ml^{-1}; Life Technologies, 16-4732-82, clone eBMG2b) or vehicle control as indicated; (3) conditioned medium from human hepatic macrophage subpopulations as indicated. The final volume was 50 μl for all conditions. Cells were then incubated in an Incucyte ZOOM live cell analysis system (Essen Biosciences) humidified at 37 °C with 5% CO_2 with imaging every 3 h using the 10× optic for either 87 h (recombinant cytokines/inhibitors) or 39 h (macrophage-conditioned medium). Analysis was performed with the Incucyte proprietary analysis software (v.2018A) by using machine learning to distinguish the individual nuclei (stained red by the NucLight Rapid Red dye) and perform nuclear counts of the images at each 3 h time point over the period of culture. Data are expressed as the AUC for percentage change in nuclear number from baseline versus time (hours), calculated in GraphPad Prism.

Gene knockdown in human HSCs

Knockdown of NOTCH3 in human HSCs was performed using siRNA. HSCs were plated at 75,000 cells per well in a 12-well plate (Costar, 3513) followed by serum starvation overnight (in HSC medium without FBS). siRNA duplexes with Lipofectamine RNAiMAX Transfection Reagent (Thermo Fisher, 13778075) were prepared in OptiMEM (Thermo Fisher, 31985070) according to the manufacturer's recommendations, and used at a concentration of 50 nM. Cells were exposed to the duplex for 48 h, in HSC medium containing 2% FBS. Cells were collected for RNA and quantitative PCR with reverse transcription (RT–qPCR). Knockdown efficiency was assessed by *NOTCH3* RT–qPCR. The best siRNA for knockdown was determined empirically using the FlexiTube GeneSolution kit (Qiagen, GS4854). HSCs treated with control siRNA (Qiagen, 1027280) and siRNA for *NOTCH3* (Qiagen, Hs_NOTCH3_3, SI00009513; knockdown 83%) were then assessed for fibrillar collagen gene expression.

Mouse HSC activation

Primary mouse HSCs were isolated from healthy mice as previously described[26]. In brief, after cannulation of the inferior vena cava, the portal vein was cut to allow retrograde step-wise perfusion with pronase (Sigma, P5147) and collagenase D (Roche, 11088882001) -containing solutions, before ex vivo digestion in a solution containing pronase, collagenase D and DNase (Roche, 10104159001). HSCs were isolated from the digest solution by Histodenz (Sigma, D2158-100G) gradient centrifugation. HSCs were plated at a density of 400,000 cells per well in a 24-well plate (Costar, 3524) in HSC medium containing 10% FBS. After overnight culture, cells were washed with PBS and cultured in HSC medium containing 2% FBS. For macrophage co-culture, transwell inserts (0.4-μm polyester membrane; Costar, 3470) were then placed above adherent HSCs. FACS-sorted $CD9^+$ mouse SAMacs or $CD9^-$ mouse TMs from CCl_4-treated mice were resuspended in HSC medium containing 2% FBS at 400,000 cells per ml and 200,000 cells were added to the top of the transwell insert. Co-culture proceeded for 48 h and HSCs were collected for RNA. Quiescent HSCs (collected at the start of co-culture) were used as a control population.

Isolation of human liver endothelial cells

Human liver endothelial cells were isolated from cirrhotic explant livers and non-fibrotic control donor liver as previously described[32]. Endothelial cells were cultured on plasticware coated with rat-tail collagen (Sigma, C3867) in complete endothelial medium consisting of endothelial basal media (Thermo Fisher, 11111044) containing 10% heat-inactivated human serum (tcsBiosciences, CS100-500), 100 U penicillin, 100 μg ml^{-1} streptomycin, 2 mM glutamine (Sigma, G6784), VEGF (10 ng ml^{-1}; Peprotech, 100-20) and 10 ng ml^{-1} HGF (10 ng ml^{-1}; Peprotech, 100-39). Expression of PLVAP, CD34, ACKR1 and JAG1 was assessed using flow cytometry.

Flow-based adhesion assays

Flow-based adhesion assays were performed as previously described[22,32]. In brief, endothelial cells from healthy and cirrhotic liver were seeded onto a rat-tail collagen-coated Ibidi slide $VI^{0.4}$ (Ibidi, 80606) at a density to give a monolayer and incubated overnight. Peripheral blood was collected from healthy donors in EDTA-coated tubes. PBMCs were isolated using a lympholyte density gradient (Cedarlane Laboratories) then washed in PBS containing 1 mM Ca^{2+}, 0.5 mM Mg^{2+} and 0.15% bovine serum albumin (BSA). Monocytes were enriched from PBMCs using a pan-monocyte isolation kit (Miltenyi Biotech, 130-096-537) according

to the manufacturer's protocol. For flow-based adhesion assays, cells were resuspended at 10^6 cells per millilitre in endothelial basal media (Thermo Fisher, 11111044) containing 0.15% BSA, then perfused over the endothelial cell monolayer for 5 min at 0.28 ml min^{-1}. Non-adherent cells were washed off during 5 min perfusion of 0.15% BSA human basal endothelial medium and 10 random non-overlapping images were randomly recorded from each channel. Total adherent (bright-phase; expressed as cell number per mm^2 per million cells perfused) and transmigrating cells (dark-phase; expressed as percentage total adherent cells) on an endothelial cell monolayer from each patient were counted and quantified as previously described[22].

Gene knockdown in endothelial cells

Knockdown of *ACKR1* and *PLVAP* gene expression in human cirrhotic endothelial cells was performed using siRNA as previously described[32]. In brief, siRNA duplexes for *PLVAP*, *ACKR1* or negative control (Qiagen, 1027280) with Lipofectamine RNAiMAX Transfection Reagent (Thermo Fisher, 13778075) were prepared in OptiMEM (Thermo Fisher, 31985070) according to the manufacturer's recommendations, and used at a concentration of 25 nM. Cells were exposed to the duplex for 4 h at 37 °C, after which time the medium was replaced with endothelial basal medium containing 10% heat-inactivated human serum for 24 h. The medium was then replaced with complete endothelial medium and incubated at 37 °C with 5% CO_2 for a further 24 h. Knockdown efficacy was assessed by flow cytometry and the MFI (Extended Data Fig. 8f). The best siRNA for knockdown was determined empirically using the FlexiTube GeneSolution kit (Qiagen, GS83483 (*PLVAP*) and GS2532 (*ACKR1*)). For flow-based adhesion assays, siRNAs against *PLVAP* (Qiagen, Hs_PLVAP_1, SI00687547; knockdown 50.6%), *ACKR1* (Qiagen, Hs_Fy_5, SI02627667; knockdown 37.7%) or control siRNA were selected. Then, 90,000 endothelial cells from cirrhotic patients ($n = 6$) were seeded into channels of a rat-tail collagen-coated Ibidi slide VI$^{0.4}$ and gene knockdown was performed, followed by flow-based adhesion assay as described above.

Co-culture of endothelial cells and HSCs

HSCs (15,000 cells) were seeded onto an Ibidi slide VI$^{0.4}$ with and without primary human endothelial cells (15,000 cells) from individual patients with cirrhosis ($n = 3$) in complete endothelial medium. After 2 h, all growth factor supplements were removed and cells were cultured for a further 72 h in endothelial basal medium containing 10% heat-inactivated human serum with or without the Notch signalling inhibitor dibenzazepine (10 μM; Bio-Techne, 4489/10) or vehicle (DMSO) control. Cells were fixed in 4% paraformaldehyde (PFA) for 30 min, permeabilized with 0.3% Triton X-100 in PBS for 5 min and blocked with 10% goat serum in PBS for 30 min followed by primary antibody incubation (mouse anti-PECAM1 and rabbit anti-collagen 1; see Supplementary Table 19) for 1 h. Cells were washed in 0.1% Triton X-100 in PBS followed by addition of fluorescently conjugated secondary antibodies (1:500 dilution) for 1 h. Cells were mounted with Pro-long Gold anti-fade DAPI, images were taken on the Confocal Microscope Zeiss LSM780, and the collagen 1 staining area was quantified using IMARIS.

RNA extraction and RT–qPCR

RNA was isolated from HSCs using the RNeasy Plus Micro Kit (Qiagen, 74034) and cDNA synthesis performed using the QuantiTect Reverse Transcription Kit (Qiagen, 205313) according to the manufacturer's protocol. Reactions were performed in triplicate in 384-well plate format and were assembled using the QIAgility automated pipetting system (Qiagen). RT–qPCR for human HSCs was performed using PowerUp SYBR Green Master Mix (Thermo Fisher, A25777) with the following primers (all Qiagen): *GAPDH* (QT00079247), *COL1A1* (QT00037793), *COL3A1* (QT00058233) and *NOTCH3* (QT00003374). RT–qPCR for mouse HSCs was performed using TaqMan Fast Advanced Master Mix (Thermo Fisher, 4444557) with the following primers: *Gapdh* (Thermo Fisher, Mm99999915_g1) and *Col3a1* (Thermo Fisher, Mm00802300_m1). Samples were amplified on an ABI 7900HT FAST PCR system (Applied Biosystems, Thermo Fisher Scientific). Data were analysed using Thermo Fisher Connect cloud qPCR analysis software (Thermo Fisher Scientific). The $2^{-\Delta\Delta Ct}$ quantification method, using *GAPDH* for normalization, was used to estimate the amount of target mRNA in samples, and expression calculated relative to average mRNA expression levels from control samples.

Immunohistochemistry, immunofluorescence and single-molecule FISH

Formalin-fixed paraffin-embedded human liver tissue was cut into 4-μm sections, dewaxed, rehydrated, then incubated in 4% neutral-buffered formalin for 20 min. After heat-mediated antigen retrieval in pH 6 sodium citrate (microwave; 15 min), slides were washed in PBS and incubated in 4% hydrogen peroxide for 10 min. Slides were then washed in PBS, blocked using protein block (GeneTex, GTX30963) for 1 h at room temperature before incubation with primary antibodies for 1 h at room temperature. A full list of primary antibodies and conditions is shown in Supplementary Table 19. Slides were washed in PBS plus 0.1% Tween 20 (PBST; Sigma-Aldrich, P1379) then incubated with ImmPress HRP Polymer Detection Reagents (depending on species of primary; rabbit, MP-7401; mouse, MP-6402-15; goat, MP-7405; all Vector Laboratories) for 30 min at room temperature. Slides were washed in PBS followed by detection. For DAB staining, sections were incubated with DAB (DAKO, K3468) for 5 min and washed in PBS before a haematoxylin (Vector Laboratories, H3404) counterstain. For multiplex immunofluorescence staining, following the incubation with ImmPress and PBS wash, initial staining was detected using Cy3, Cy5, or fluorescein tyramide (Perkin-Elmer, NEL741B001KT) at a 1:1,000 dilution. Slides were then washed in PBST followed by further heat treatment with pH 6 sodium citrate (15 min), washes in PBS, protein block, incubation with the second primary antibody (incubated overnight at 4 °C), ImmPress Polymer and tyramide as before. This sequence was repeated for the third primary antibody (incubated at room temperature for 1 h) and a DAPI-containing mountant was then applied (Thermo Fisher Scientific, P36931).

For AMEC staining (only CLEC4M immunohistochemistry), all washes were carried out with TBST (dH_2O, 200 mM Tris, 1.5 M NaCl, 1% Tween-20 (all Sigma-Aldrich) pH 7.5) and peroxidase blocking was carried out for 30 min in 0.6% hydrogen peroxide in methanol. Sections were incubated with AMEC (Vector Laboratories, SK-4285) for 20 min and washed in TBST before a haematoxylin (Vector Laboratories, SK-4285) counterstain.

For combined single-molecule fluorescent in situ hybridization (smFISH) and immunofluorescence, detection of TREM2 was performed using the RNAscope 2.5 LS Reagent Kit BrownAssay (Advanced Cell Diagnostics) in accordance with the manufacturer's instructions. In brief, 5-μm tissue sections were dewaxed, incubated with endogenous enzyme block, boiled in pre-treatment buffer and treated with protease, followed by target probe hybridization using the RNAscope LS 2.5 Hs-TREM2 (420498, Advanced Cell Diagnostics) probe. Target RNA was then detected with Cy3 tyramide (Perkin-Elmer, NEL744B001KT) at a 1:1,000 dilution. The sections were then processed through a pH 6 sodium citrate heat-mediated antigen retrieval, hydrogen peroxidase treatment and protein block (all as for multiplex immunofluorescence staining as above). MNDA antibody was applied overnight at 4 °C, completed using a secondary ImmPress HRP Anti-Rabbit Peroxidase IgG (Vector Laboratories, MP7401), visualized using a fluorescein tyramide (Perkin-Elmer, NEL741B001KT) at a 1:1,000 dilution and stained with DAPI.

Bright-field and fluorescently stained sections were imaged using the slide scanner AxioScan.Z1 (Zeiss) at 20× magnification (40× magnification for smFISH). Images were processed and scale bars added using Zen Blue (Zeiss) and Fiji software[33].

Cell counting and image analysis

Automated cell counting was performed using QuPath software[34]. In brief, DAB-stained whole tissue section slide-scanned images (CZI files)

were imported into QuPath. Cell counts were carried out using the positive cell detection tool, detecting haematoxylin-stained nuclei and then thresholding for positively stained DAB cells, generating DAB-positive cell counts per mm^2 tissue. Identical settings and thresholds were applied to all slides for a given stain and experiment. For cell counts of fibrotic septae versus parenchymal nodules, the QuPath segmentation tool was used to segment the DAB-stained whole tissue section into fibrotic septae or non-fibrotic parenchymal nodule regions using tissue morphological characteristics (Fig. 2j). Positive cell detection was then applied to the fibrotic and non-fibrotic regions in turn, providing DAB-positive cell counts per mm^2 in fibrotic septae and non-fibrotic parenchymal nodules for each tissue section.

Digital morphometric pixel analysis was performed using the Trainable Weka Segmentation (TWS) plugin[35] in Fiji software[33]. In brief, each stained whole tissue section slide-scanned image was converted into multiple TIFF files in Zen Blue software (Zeiss). TIFF files were imported into Fiji and TWS plugin trained to produce a classifier which segments images into areas of positive staining, tissue background and white space[35]. The same trained classifier was then applied to all TIFF images from every tissue section for a particular stain, providing a percentage area of positive staining for each tissue section. For digital morphometric quantification of positive staining of fibrotic septae versus parenchymal nodules, TIFF images were segmented into fibrotic septae or non-fibrotic parenchymal nodule regions using tissue morphological characteristics, followed by analysis using the TWS plugin in Fiji software.

Histological assessment of NASH sections

Sections stained with haematoxylin and eosin or picrosirius red were whole-slide imaged using a NanoZoomer imager (Hamamatsu Photonics). Images of stained sections were independently scored by a consultant liver transplant histopathologist (T.J.K.) at the national liver transplant centre with experience in trial scoring by applying the ordinal NAFLD activity score[36]. For observer-independent quantification of the area of positive picrosirius red staining, images were split using ndpisplit[37] into tiles of ×5 magnification before the application of a classifier that had been trained by the liver histopathologist using the machine learning WEKA plugin in FIJI[33,35], as previously described[38]. All analysis was undertaken blinded to all other data.

Droplet-based scRNA-seq

Single cells were processed through the Chromium Single Cell Platform using the Chromium Single Cell 3′ Library and Gel Bead Kit v2 (10X Genomics, PN-120237) and the Chromium Single Cell A Chip Kit (10X Genomics, PN-120236) as per the manufacturer's protocol. In brief, single cells were sorted into PBS plus 0.1% BSA, washed twice and counted using a Bio-Rad TC20. Then, 10,800 cells were added to each lane of the 10X chip. The cells were partitioned into Gel Beads in Emulsion in the Chromium instrument, in which cell lysis and bar-coded reverse transcription of RNA occurred, followed by amplification, fragmentation and 5′ adaptor and sample index attachment. Libraries were sequenced on an Illumina HiSeq 4000.

Computational analysis

In total, we analysed 67,494 human cells from healthy ($n = 5$) and cirrhotic ($n = 5$) livers, 30,741 PBMCs from patients with cirrhosis ($n = 4$) and compared our data with a publicly available reference dataset of 8,381 PBMCs from a healthy donor (https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc8k).

Pre-processing scRNA-seq data

We aligned to the GRCh38 and mm10 (Ensembl 84) reference genomes as appropriate for the input dataset, and estimated cell-containing partitions and associated unique molecular identifiers (UMIs), using the Cell Ranger v.2.1.0 Single-Cell Software Suite from 10X Genomics. Genes expressed in fewer than three cells in a sample were excluded, as were cells that expressed fewer than 300 genes or mitochondrial gene content >30% of the total UMI count. We normalized by dividing the UMI count per gene by the total UMI count in the corresponding cell and log-transforming. Variation in UMI counts between cells was regressed according to a negative binomial model, before scaling and centring the resulting value by subtracting the mean expression of each gene and dividing by its standard deviation (E_n), then calculating $\ln(10^4 \times E_n + 1)$.

Dimensionality reduction, clustering and differential expression analysis

We performed unsupervised clustering and differential gene expression analyses in the Seurat R package v.2.3.0[39]. In particular, we used shared nearest neighbour graph-based clustering, in which the graph was constructed using from 2 to 11 principal components as determined by dataset variability shown in principal component analysis (PCA); the resolution parameter to determine the resulting number of clusters was also tuned accordingly. To assess cluster similarity we used the 'BuildClusterTree' function from Seurat.

In total, we present scRNA-seq data from ten human liver samples (named healthy 1–5 and cirrhotic 1–5), five human blood samples ($n = 4$ cirrhotic named blood 1–4 and $n = 1$ healthy named PBMC8K; pbmc8k dataset sourced from single-cell gene expression datasets hosted by 10X Genomics), and two mouse liver samples ($n = 3$ uninjured and $n = 3$ fibrotic). For seven human liver samples (healthy 1–4 and cirrhotic 1–3), we performed scRNA-seq on both leucocytes ($CD45^+$) and other non-parenchymal cells ($CD45^-$); for the remaining three human livers (healthy 5, cirrhotic 4–5) we performed scRNA-seq on leucocytes only (Extended Data Fig. 2e, f).

Initially, we combined all human scRNA-seq datasets (liver and blood) and performed clustering analysis with the aim of isolating a population of liver-resident cells, by identifying contaminating circulatory cells within datasets generated from liver digests and removing them from downstream analysis. Specifically, we removed from our liver datasets cells that fell into clusters 1 and 13 of the initial dataset in Extended Data Fig. 1d.

Using further clustering followed by signature analysis, we interrogated this post-processed liver-resident cell dataset for robust cell lineages. These lineages were isolated into individual datasets, and the process was iterated to identify robust lineage subpopulations. At each stage of this process we removed clusters expressing more than one unique lineage signature in more than 25% of their cells from the dataset as probable doublets. This resulted in removal of 1,351 cells. Where the cell proliferation signature identified distinct cycling subpopulations, we re-clustered these again to ascertain the identity of their constituent cells.

The mouse scRNA-seq datasets were combined, clustered and interrogated for cell lineages in a similar manner to their human counterparts.

All heat maps, *t*-distributed stochastic neighbour embedding (*t*-SNE) and UMAP visualizations, violin plots and dot plots were produced using Seurat functions in conjunction with the ggplot2, pheatmap and grid R packages. *t*-SNE and UMAP visualizations were constructed using the same number of principal components as the associated clustering, with perplexity ranging from 30 to 300 according to the number of cells in the dataset or lineage. We conducted differential gene expression analysis in Seurat using the standard AUC classifier to assess significance. We retained only those genes with a log-transformed fold change of at least 0.25 and expression in at least 25% of cells in the cluster under comparison.

Defining cell lineage signatures

For each cell, we obtained a signature score across a curated list of known marker genes per cell lineage in the liver (Supplementary Table 2). This signature score was defined as the geometric mean of the expression of the associated signature genes in that cell. Lineage signature scores were scaled from 0 to 1 across the dataset, and the score for each cell

with a signature less than a given threshold (the mean of said signature score across the entire dataset) was set to 0.

Batch effect and quality control

To investigate agreement between samples, we extracted the average expression profile for a given cell lineage in each sample, and calculated the Pearson correlation coefficients between all possible pairwise comparisons of samples per lineage[40].

Imputing dropout in T cell and ILC clusters

To impute dropout of low-abundance transcripts in our T cell and ILC clusters so that we might associate them with known subpopulations, we downsampled to 7,380 cells from 36,900 and applied the scImpute R package v.0.0.8[41], using as input both our previous annotation labels and *k*-means spectral clustering ($k = 5$), but otherwise default parameters.

Analysing functional phenotypes of scar-associated cells

For further analysis of function we adopted the self-organizing maps approach as implemented in the SCRAT R package v.1.0.0[42]. For each lineage of interest, we constructed a self-organizing map in SCRAT using default input parameters and according to its clusters. We defined the signatures expressed in a cell by applying a threshold criterion ($e^{thresh} = 0.95 \times e^{max}$) selecting the highest-expressed metagenes in each cell, and identified for further analysis those metagene signatures defining at least 30% of cells in at least one cluster within the lineage. We smoothed these self-organizing maps using the 'disaggregate' function from the raster R package for visualization purposes, and scaled radar plots to maximum proportional expression of the signature. Gene Ontology enrichment analysis on the genes in these spots was performed using PANTHER 13.1 (pantherdb.org).

Inferring injury dynamics and transcriptional regulation

To generate cellular trajectories (pseudotemporal dynamics) we used the monocle R package v.2.6.1[43]. We ordered cells in a semi-supervised manner on the basis of their Seurat clustering, scaled the resulting pseudotime values from 0 to 1, and mapped them onto either the *t*-SNE or UMAP visualizations generated by Seurat or diffusion maps as implemented in the scater R package v.1.4.0[44] using the top 500 variable genes as input. We removed mitochondrial and ribosomal genes from the gene set for the purposes of trajectory analysis. Differentially expressed genes along this trajectory were identified using generalized linear models via the 'differentialGeneTest' function in monocle.

When determining significance for differential gene expression along the trajectory, we set a *q*-value threshold of 1×10^{-20}. We clustered these genes using hierarchical clustering in pheatmap, cutting the tree at $k = 3$ to obtain gene modules with correlated gene expression across pseudotime. Cubic smoothing spline curves were fitted to scaled gene expression along this trajectory using the smooth.spline command from the stats R package, and Gene Ontology enrichment analysis again performed using PANTHER 13.1.

We verified the trajectory and its directionality using the velocyto R package v.0.6.0[17], estimating cell velocities from their spliced and unspliced mRNA content. We generated annotated spliced and unspliced reads from the 10X BAM files via the 'dropEst' pipeline, before calculating gene-relative velocity using *k*NN pooling with $k = 25$, determining slope gamma with the entire range of cellular expression, and fitting gene offsets using spanning reads. Aggregate velocity fields (using Gaussian smoothing on a regular grid) and transition probabilities per lineage subpopulations were visualized on *t*-SNE, UMAP, or diffusion map visualizations as previously generated. Gene-specific phase portraits were plotted by calculating spliced and unspliced mRNA levels against steady-state inferred by a linear model; levels of unspliced mRNA above and below this steady-state indicate increasing and decreasing expression of said gene, respectively. Similarly, we plotted the unspliced count signal residual per gene, based on the estimated gamma fit, with positive and negative residuals indicating expected upregulation and downregulation, respectively.

For transcription factor analysis, we obtained a list of all genes identified as acting as transcription factors in humans from AnimalTFDB[45]. To analyse transcription factor regulons further, we adopted the SCENIC v.0.1.7 workflow in R[46], using default parameters and the normalized data matrices from Seurat as input. For visualization, we mapped the regulon activity (AUC) scores thus generated to the pseudotemporal trajectories from monocle and the clustering subpopulations from Seurat.

Analysing inter-lineage interactions within the fibrotic niche

For comprehensive systematic analysis of inter-lineage interactions within the fibrotic niche, we used CellPhoneDB[27]. CellPhoneDB is a manually curated repository of ligands, receptors and their interactions, integrated with a statistical framework for inferring cell–cell communication networks from single-cell transcriptomic data. In brief, we derived potential ligand–receptor interactions on the basis of the expression of a receptor by one lineage subpopulation and a ligand by another; as input to this algorithm, we used cells from the fibrotic niche as well as liver sinusoidal endothelial cells and KCs as controls, and we considered only ligands and receptors expressed in greater than 5% of the cells in any given subpopulation. Subpopulation-specific interactions were identified as follows: (1) randomly permuting the cluster labels of all cells 1,000 times and determining the mean of the average receptor expression of a subpopulation and the average ligand expression of the interacting subpopulation, thus generating a null distribution for each ligand–receptor pair in each pairwise comparison between subpopulations; (2) calculating the proportion of these means that were 'as or more extreme' than the actual mean, thus obtaining a *P* value for the likelihood of subpopulation specificity for a given ligand–receptor pair; (3) prioritizing interactions that displayed specificity to subpopulations interacting within the fibrotic niche.

Canonical correlation analysis

To compare human and mouse populations of monocytic phagocytes, we used canonical correlation analysis as implemented in Seurat[18]. We map the genes in the human dataset to their mouse orthologues using biomaRt, discarding any genes for which no orthologues can be found. We then calculate the shared low-dimensional subspace on the union of genes that are variably expressed in both datasets ($n = 159$), and align using six canonical components as determined by evaluating the biweight midcorrelation. Results are visualized by *t*-SNE analysis as previously described.

Deconvolution of whole liver microarray data

To assess the macrophage composition of early-stage NAFLD, we performed deconvolution analysis on publicly available microarray data from annotated liver biopsy specimens taken across the NAFLD disease spectrum (GEO accession GSE48452)[20]. Tissue MP cells from our human scRNA-seq data were manually clustered into the main annotated MP populations. Signature gene expression profiles of SAMacs, TMs and KCs were used to deconvolve the monocyte–macrophage composition of liver biopsy samples from GSE48452 using Cibersort[47], as previously described[19]. The monocyte–macrophage composition of each biopsy sample was then compared to the associated histological and demographic features, available from the GEO database.

Statistics and reproducibility

To assess whether our identified subpopulations were significantly overexpressed in injury, we posited the proportion of injured cells in each cluster as a random count variable using a Poisson process, as previously described[40]. We modelled the rate of detection using the total number of cells in the lineage profiled in a given sample as an offset, with the condition of each sample (healthy versus cirrhotic) provided as a

covariate factor. The model was fitted using the R command 'glm' from the stats package. The *P* value for the significance of the proportion of injured cells was assessed using a Wald test on the regression coefficient.

Remaining statistical analyses were performed using GraphPad Prism. Comparison of changes between two groups was performed using a Mann–Whitney test (unpaired; two-tailed) or a Wilcoxon matched-pairs signed rank test (paired; two-tailed). Comparison of changes between multiple groups was performed using a Kruskal–Wallis and Dunn, one-way ANOVA and Tukey or repeated measures one-way ANOVA and Tukey tests. Correlations were preformed using Pearson correlation and best-fit line plotted using linear regression. $P < 0.05$ was considered statistically significant. All immunofluorescence stains were repeated in a minimum of three patients and representative images are displayed.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Our expression data are freely available for user-friendly interactive browsing online at http://www.livercellatlas.mvm.ed.ac.uk. CellPhoneDB is available at www.CellPhoneDB.org. All raw sequencing data have been deposited in the Gene Expression Omnibus (GEO) under accession GSE136103.

Code availability

R scripts enabling the main steps of the analysis are available from the corresponding authors on reasonable request.

29. Searle, B. C., Gittelman, R. M., Manor, O. & Akey, J. M. Detecting sources of transcriptional heterogeneity in large-scale RNA-seq data sets. *Genetics* **204**, 1391–1396 (2016).
30. Bain, C. C. et al. Long-lived self-renewing bone marrow-derived macrophages displace embryo-derived cells to inhabit adult serous cavities. *Nat. Commun.* **7**, 11852 (2016).
31. Heinrich, M. C. et al. Crenolanib inhibits the drug-resistant PDGFRA D842V mutation associated with imatinib-resistant gastrointestinal stromal tumors. *Clin. Cancer Res.* **18**, 4375–4384 (2012).
32. Patten, D. A. et al. SCARF-1 promotes adhesion of $CD4^+$ T cells to human hepatic sinusoidal endothelium under conditions of shear stress. *Sci. Rep.* **7**, 17600 (2017).
33. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
34. Bankhead, P. et al. QuPath: open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878 (2017).
35. Arganda-Carreras, I. et al. Trainable Weka Segmentation: a machine learning tool for microscopy pixel classification. *Bioinformatics* **33**, 2424–2426 (2017).
36. Kleiner, D. E. et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology* **41**, 1313–1321 (2005).
37. Deroulers, C. et al. Analyzing huge pathology images with open source software. *Diagn. Pathol.* **8**, 92 (2013).
38. Kendall, T. J. et al. Hepatic elastin content is predictive of adverse outcome in advanced fibrotic liver disease. *Histopathology* **73**, 90–100 (2018).
39. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
40. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
41. Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* **9**, 997 (2018).
42. Camp, J. G. et al. Multilineage communication regulates human liver bud development from pluripotency. *Nature* **546**, 533–538 (2017).
43. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
44. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).
45. Zhang, H. M. et al. AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res.* **43**, D76–D81 (2015).
46. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
47. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).

Acknowledgements This work was supported by an MRC Clinician Scientist Fellowship (MR/N008340/1) to P.R., a Wellcome Trust Senior Research Fellowship in Clinical Science (ref. 103749) to N.C.H., an AbbVie Future Therapeutics and Technologies Division grant to N.C.H., a Guts UK–Children's Liver Disease Foundation grant (ref. R43927) to N.C.H., a Tenovus Scotland grant (ref. E18/05) to R.D. and N.C.H. and British Heart Foundation grants (RM/17/3/33381; RE/18/5/34216) to N.C.H. R.V.-T. was funded by EMBO and Human Frontiers long-term fellowships. C.J.W. was funded by a BBSRC New Investigator Award (BB/N018869/1). P.N.N., C.J.W. and N.T.L. are funded by the NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. This paper presents independent research supported by the NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. J.P.I. is funded by the NIHR Bristol Biomedical Research Centre, University Hospitals Bristol Foundation Trust and the University of Bristol. C.P.P. was funded by the UK Medical Research Council, MC_UU_00007/15. This work was also supported by Wellcome Sanger core funding (WT206194). We thank the patients who donated liver tissue and blood for this study. We thank J. Davidson, C. Ibbotson, J. Black and A. Baird of the Scottish Liver Transplant Unit and the research nurses of the Wellcome Trust Clinical Research Facility for assistance with consenting patients for this study. We thank the liver transplant coordinators and surgeons of the Scottish Liver Transplant Unit and the surgeons and staff of the Hepatobiliary Surgical Unit, Royal Infirmary of Edinburgh for assistance in procuring human liver samples. We thank S. Johnston, W. Ramsay and M. Pattison for technical assistance with FACS and flow cytometry. We thank J. Henderson for technical support and G. Muirhead for assistance with isolation of liver endothelial cells. This publication is part of the Human Cell Atlas (www.humancellatlas.org/publications).

Author contributions P.R. performed experimental design, tissue procurement, data generation, data analysis and interpretation, and manuscript preparation; R.D. performed experimental design, data generation and data analysis; E.F.D., K.P.M., B.E.P.H., M.B., J.A.M. and N.T.L. performed data generation and analysis; J.R.P. generated the interactive online browser; M.E. and R.V.-T. assisted with CellPhoneDB analyses and critically appraised the manuscript; T.J.K. performed pathological assessments and provided intellectual contribution; N.O.C., J.A.F. and P.N.N. provided intellectual contribution; C.J.W. performed tissue procurement, data generation, interpretation and intellectual contribution; J.R.W.-K. performed computational analysis with assistance from J.R.P. and R.S.T. and advice from C.P.P., J.C.M. and S.A.T.; J.R.W.-K. also helped with manuscript preparation, and C.P.P., J.C.M. and S.A.T. critically appraised the manuscript; E.M.H., D.J.M. and S.J.W. procured human liver tissue and critically appraised the manuscript. J.P.I., F.T. and J.W.P. provided intellectual contribution and critically appraised the manuscript; N.C.H. conceived the study, designed experiments, interpreted data and prepared the manuscript.

Competing interests The authors declare no competing interests.

Additional information

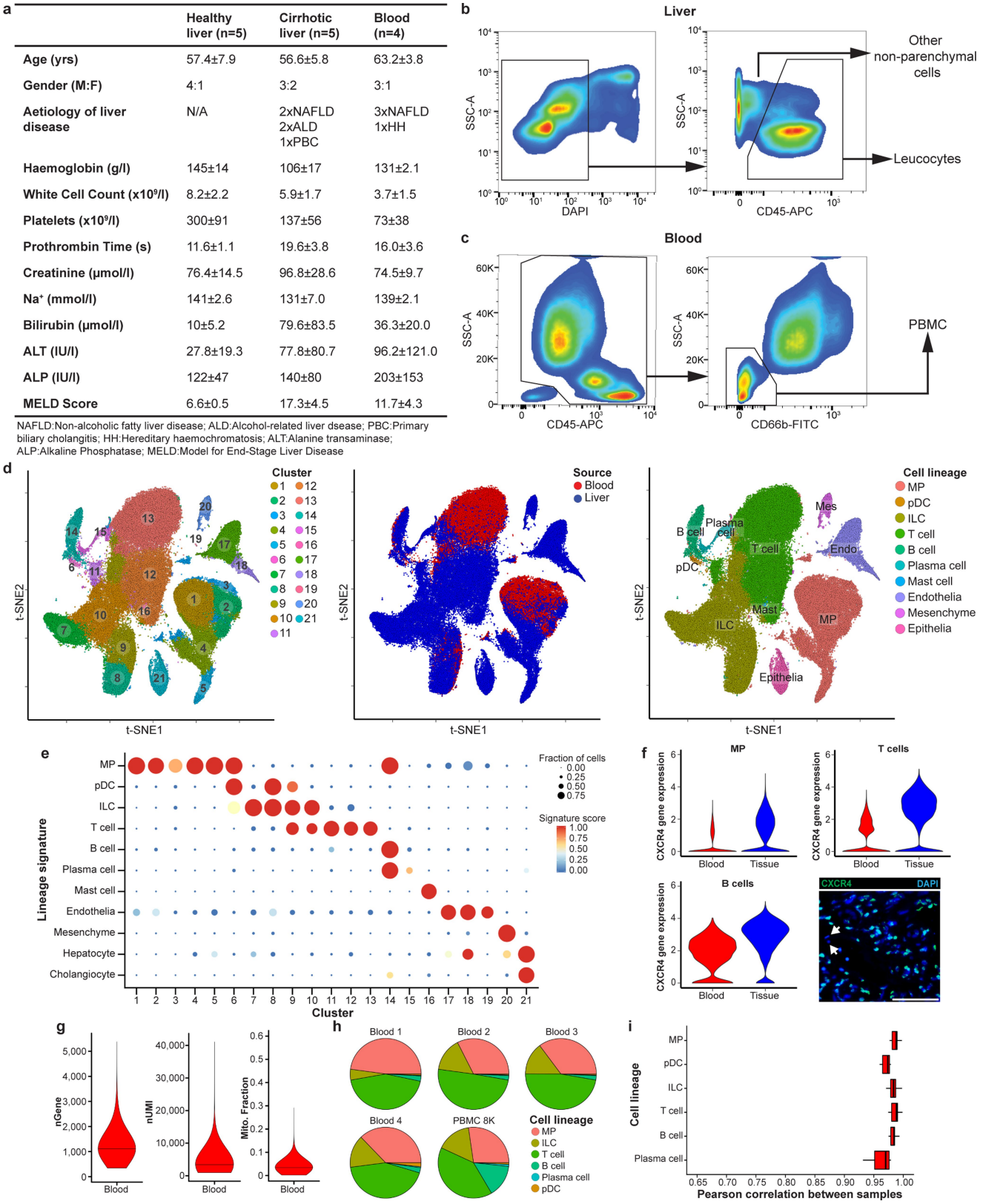
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1631-3.

Correspondence and requests for materials should be addressed to P.R. or N.C.H.

a

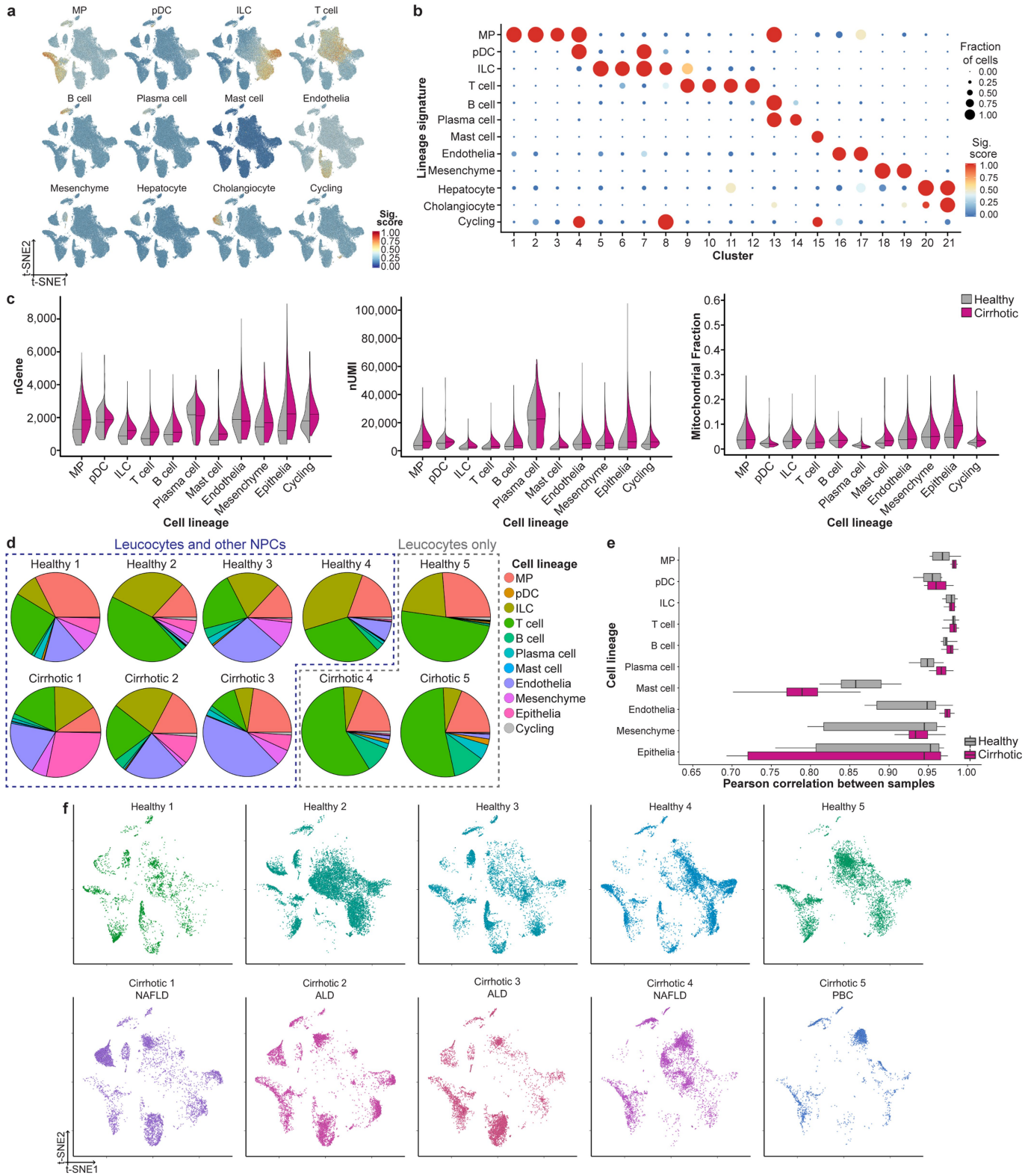
	Healthy liver (n=5)	Cirrhotic liver (n=5)	Blood (n=4)
Age (yrs)	57.4±7.9	56.6±5.8	63.2±3.8
Gender (M:F)	4:1	3:2	3:1
Aetiology of liver disease	N/A	2xNAFLD 2xALD 1xPBC	3xNAFLD 1xHH
Haemoglobin (g/l)	145±14	106±17	131±2.1
White Cell Count (x10⁹/l)	8.2±2.2	5.9±1.7	3.7±1.5
Platelets (x10⁹/l)	300±91	137±56	73±38
Prothrombin Time (s)	11.6±1.1	19.6±3.8	16.0±3.6
Creatinine (μmol/l)	76.4±14.5	96.8±28.6	74.5±9.7
Na⁺ (mmol/l)	141±2.6	131±7.0	139±2.1
Bilirubin (μmol/l)	10±5.2	79.6±83.5	36.3±20.0
ALT (IU/l)	27.8±19.3	77.8±80.7	96.2±121.0
ALP (IU/l)	122±47	140±80	203±153
MELD Score	6.6±0.5	17.3±4.5	11.7±4.3

NAFLD:Non-alcoholic fatty liver disease; ALD:Alcohol-related liver dsease; PBC:Primary biliary cholangitis; HH:Hereditary haemochromatosis; ALT:Alanine transaminase; ALP:Alkaline Phosphatase; MELD:Model for End-Stage Liver Disease

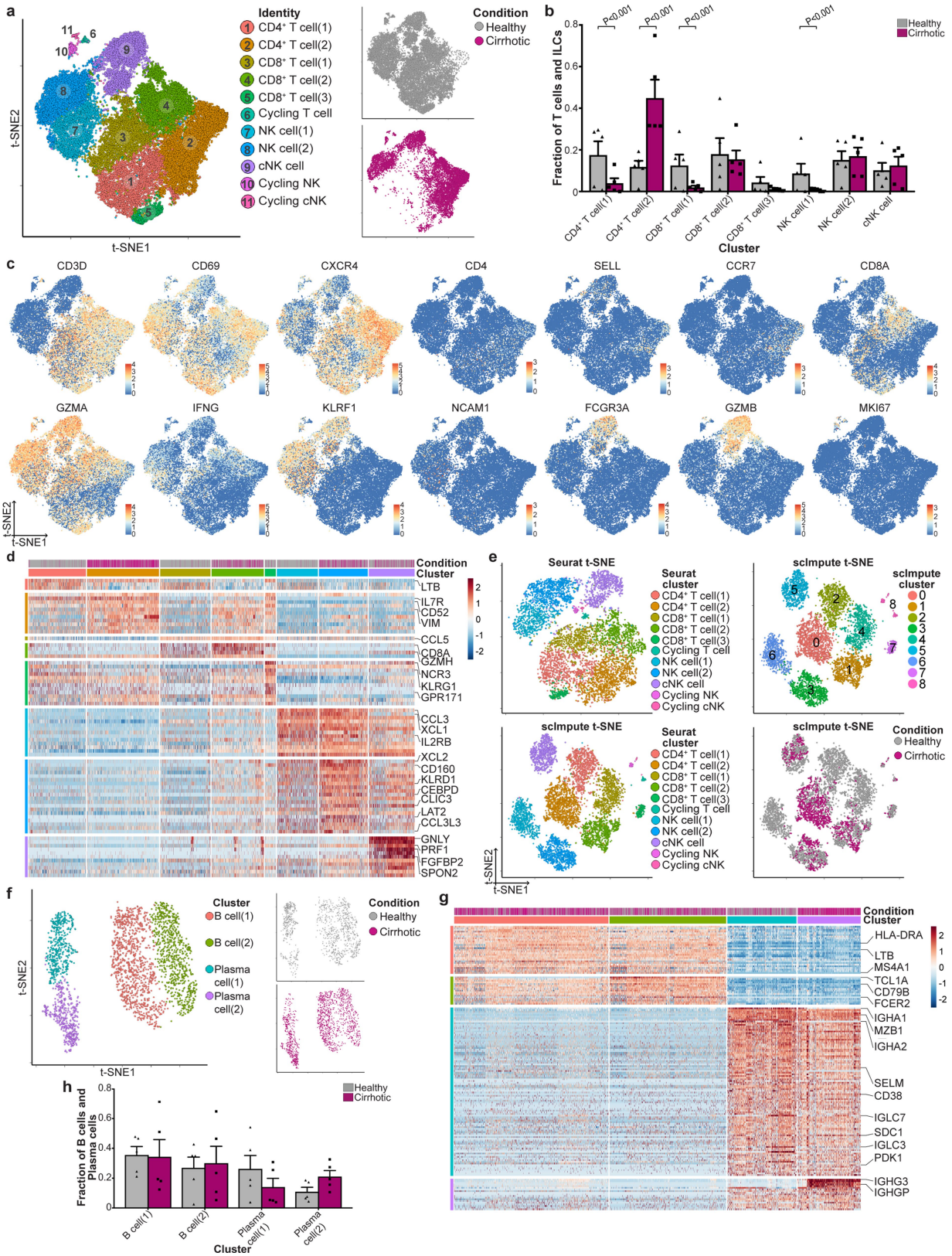


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Strategy for isolation of human liver non-parenchymal cells. a, Patient demographics and clinical information. Data are mean ± s.e.m. **b**, Flow cytometry gating strategy for isolation of leucocytes ($CD45^+$) and other non-parenchymal cells ($CD45^-$) from human liver; representative plots from ten livers. **c**, Flow cytometry gating strategy for isolation of PBMCs; representative plots from four patients. **d**, Clustering 103,568 cells from healthy ($n = 5$) and cirrhotic ($n = 5$) livers, healthy PBMCs ($n = 1$) and cirrhotic PBMCs ($n = 4$) (left), annotating the source (PBMC versus liver; middle) and cell lineage inferred from known marker gene signatures (right). **e**, Dot plot annotating PBMC and liver clusters by lineage signatures. Circle size indicates cell fraction expressing signature greater than mean; colour indicates mean signature expression (red, high; blue, low). **f**, *CXCR4* gene expression in single cells derived from blood or liver tissue, divided by cell lineage. Bottom right, representative immunofluorescence image ($n \geq 3$) of CXCR4 (green) and DAPI (blue) in human liver; arrows denote $CXCR4^-$ cells in the lumen of a blood vessel. Scale bar, 50 µm. **g**, Violin plots showing the number of unique genes (nGene), number of total unique molecular identifiers (nUMI) and mitochondrial gene fraction expressed in five PBMC samples. Black lines denote the median. **h**, Pie charts showing the proportion of cell lineages per PBMC sample. **i**, Box and whisker plots showing the agreement in expression profiles across five PBMC samples. Pearson correlation coefficients between average expression profiles for cells in each lineage, across all pairs of samples. Black bars denote the median; box edges denote the twenty-fifth and seventy-fifth percentiles; whiskers denote the full range.

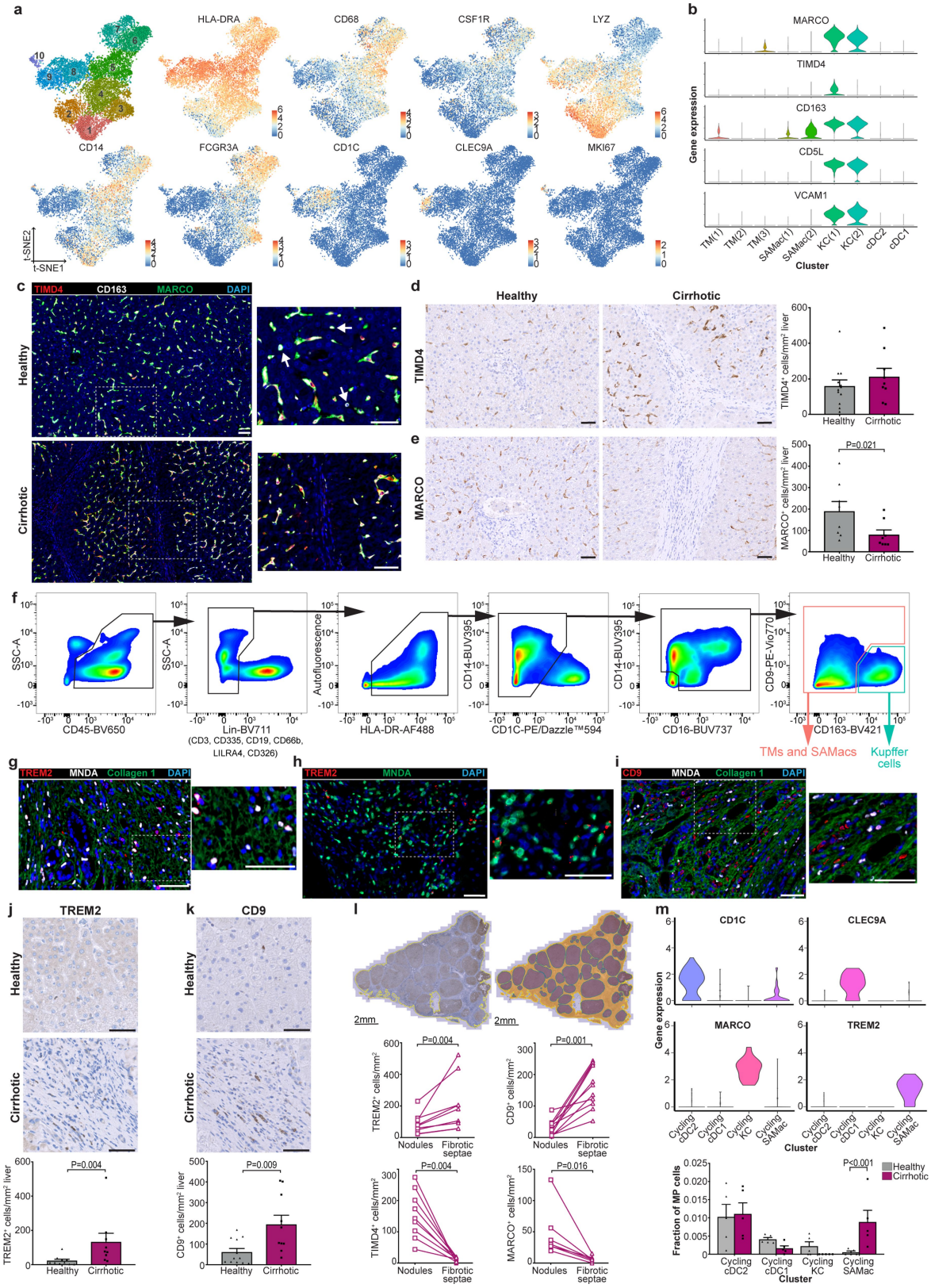


Extended Data Fig. 2 | Quality control and annotation of human liver-resident cells. a, Lineage signature expression across 66,135 liver-resident cells from healthy ($n = 5$) and cirrhotic ($n = 5$) human livers (red, high; blue, low). **b**, Dot plot annotating liver-resident cell clusters by lineage signature. Circle size indicates cell fraction expressing signature greater than mean; colour indicates mean signature expression (red, high; blue, low). **c**, Violin plots of the number of unique genes (left), number of total UMIs (middle) and mitochondrial gene fraction (right) across 66,135 liver-resident cells from healthy ($n = 5$) and cirrhotic ($n = 5$) livers. Black lines denote the median. **d**, Pie charts of the proportion of cell lineage per liver sample. **e**, Box and whisker plots of the agreement in expression profiles across healthy ($n = 5$) and cirrhotic ($n = 5$) liver samples, as in Extended Data Fig. 1i. **f**, *t*-SNE visualization of liver-resident cells per liver sample, with cirrhotic samples annotated by aetiology of underlying liver disease. ALD, alcohol-related liver disease; PBC, primary biliary cholangitis.



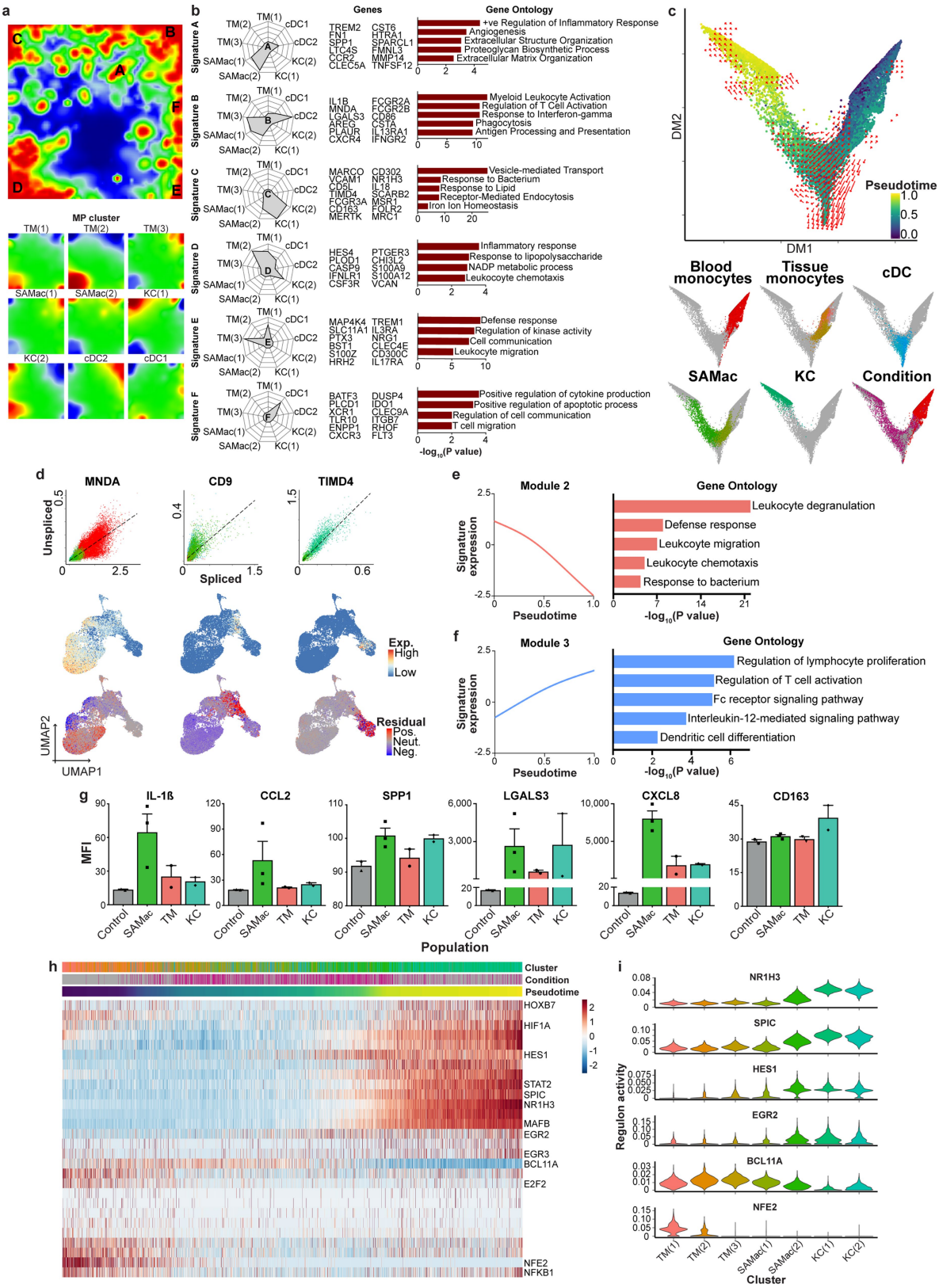
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Annotating human liver lymphoid cells. a, Clustering of 36,900 T cells and ILCs (left) from healthy ($n = 5$) and cirrhotic ($n = 5$) human livers, annotating the injury condition (right). NK, natural killer cell; cNK, cytotoxic NK cell. **b**, Fractions of T cell and ILC subpopulations in healthy ($n = 5$) and cirrhotic ($n = 5$) livers. **c**, Selected gene expression in 36,900 T cells and ILCs. **d**, Heat map of T cell and ILC cluster marker genes (colour-coded by cluster and condition), with exemplar genes labelled (right). Columns denote cells; rows denote genes. **e**, *t*-SNE visualizations of downsampled T cell and ILC dataset (7,380 cells from healthy ($n = 5$) and cirrhotic ($n = 5$) human livers) before and after imputation (scImpute); annotating data used for visualization and clustering, inferred lineage and injury condition. No additional heterogeneity was observed after imputation. **f**, Clustering 2,746 B cells and plasma cells (left) from healthy ($n = 5$) and cirrhotic ($n = 5$) human livers, annotating the injury condition (right). **g**, Heat map of B cell and plasma cell cluster marker genes (colour-coded by cluster and condition), with exemplar genes labelled (right). Columns denote cells; rows denote genes. **h**, Fractions of B cell and plasma cell subpopulations in healthy ($n = 5$) and cirrhotic ($n = 5$) livers. Data are mean ± s.e.m. *P* values determined by Wald test (**b**).



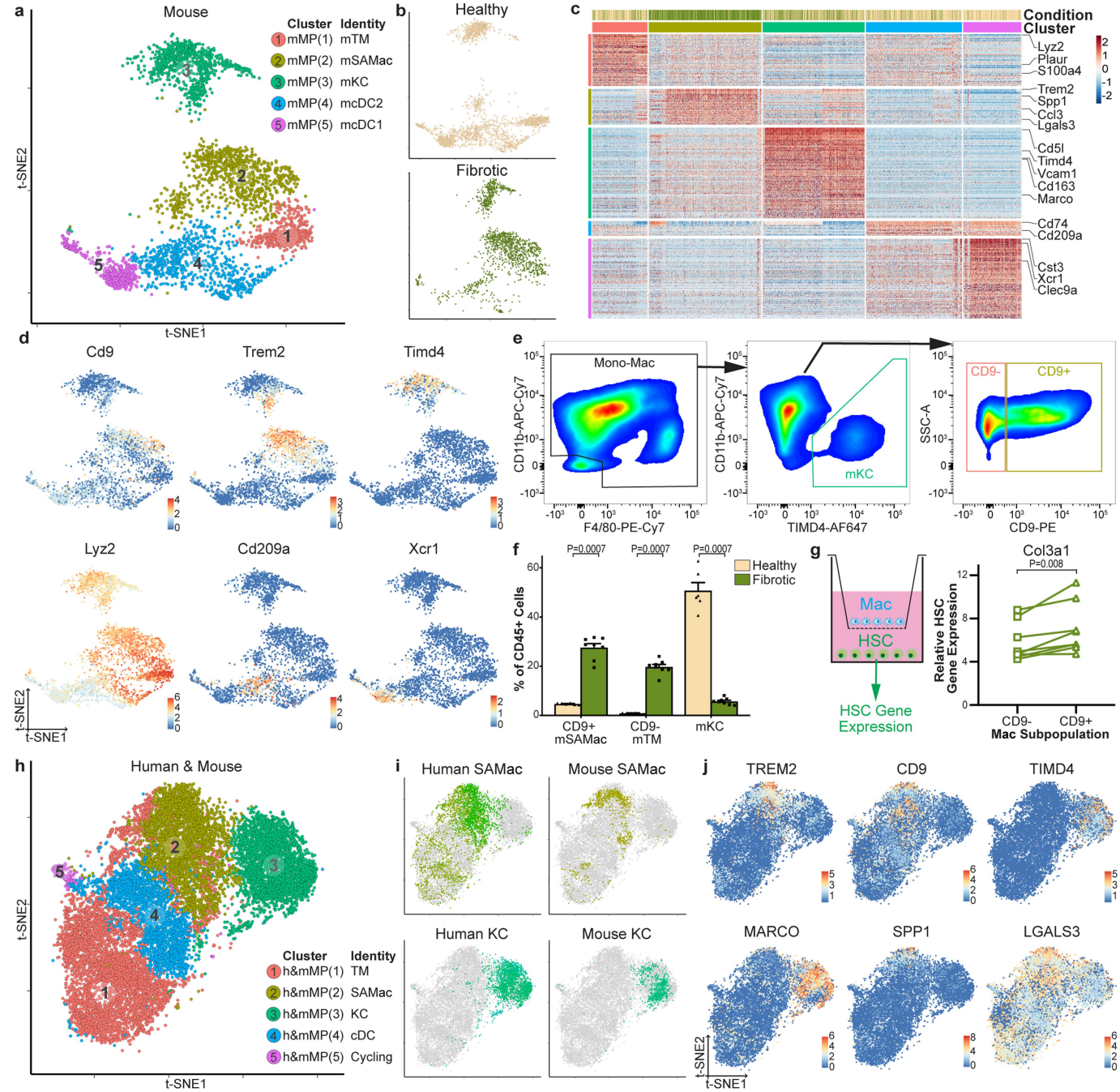
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Annotating human liver MPs. a, Clustering and selected genes expressed in 10,737 MPs from healthy ($n = 5$) and cirrhotic ($n = 5$) human livers. **b**, Scaled gene expression of KC cluster markers across MP cells from healthy ($n = 5$) and cirrhotic ($n = 5$) livers. **c**, Representative immunofluorescence images ($n \geq 3$) of TIMD4 (red), CD163 (white), MARCO (green) and DAPI (blue) in healthy and cirrhotic liver; arrows denote $CD163^{+}MARCO^{+}TIMD4^{-}$ cells.
d, Immunohistochemistry (left) and cell counts (right) of TIMD4 expression in healthy ($n = 12$) and cirrhotic ($n = 9$) human liver. **e**, Immunohistochemistry (left) and cell counts (right) of MARCO expression in healthy ($n = 8$) and cirrhotic ($n = 8$) liver. **f**, Flow cytometry gating strategy for identifying KCs, TMs and SAMacs in healthy ($n = 2$) and cirrhotic ($n = 3$) liver. SAMacs are detected as $TREM2^{+}CD9^{+}$ cells within the TM and SAMac gate (see Fig. 2f). **g**, Representative immunofluorescence images ($n \geq 3$) of TREM2 (red), MNDA (white), collagen 1 (green) and DAPI (blue) in cirrhotic liver. **h**, Representative images ($n = 2$) of TREM2 (smFISH; red), MNDA (immunofluorescence; green) and DAPI (blue) in cirrhotic liver. **i**, Representative immunofluorescence images ($n \geq 3$) of CD9 (red), MNDA (white), collagen 1 (green) and DAPI (blue) in cirrhotic liver. **j**, Immunohistochemistry (top) and cell counts (bottom) of TREM2 expression in healthy ($n = 10$) and cirrhotic ($n = 9$) liver. **k**, Immunohistochemistry (top) and cell counts (bottom) of CD9 expression in healthy ($n = 12$) and cirrhotic ($n = 10$) liver. **l**, Top, exemplar tissue segmentation of cirrhotic liver section into fibrotic septae (orange) and parenchymal nodules (purple). Bottom, cell counts based on immunohistochemistry analysis of TREM2 ($n = 9$), CD9 ($n = 11$), TIMD4 ($n = 9$) and MARCO ($n = 7$) in parenchymal nodules and fibrotic septae. **m**, Top, clustering and annotation of 208 cycling MP cells from healthy ($n = 5$) and cirrhotic ($n = 5$) livers, with scaled gene expression of MP subpopulation markers across four clusters of cycling MP cells. Bottom, fractions of cycling MP subpopulations in healthy ($n = 5$) and cirrhotic ($n = 5$) livers. All scale bars, 50 μm. Data are mean ± s.e.m. *P* values determined by two-tailed Mann–Whitney (**e**, **j**, **k**),

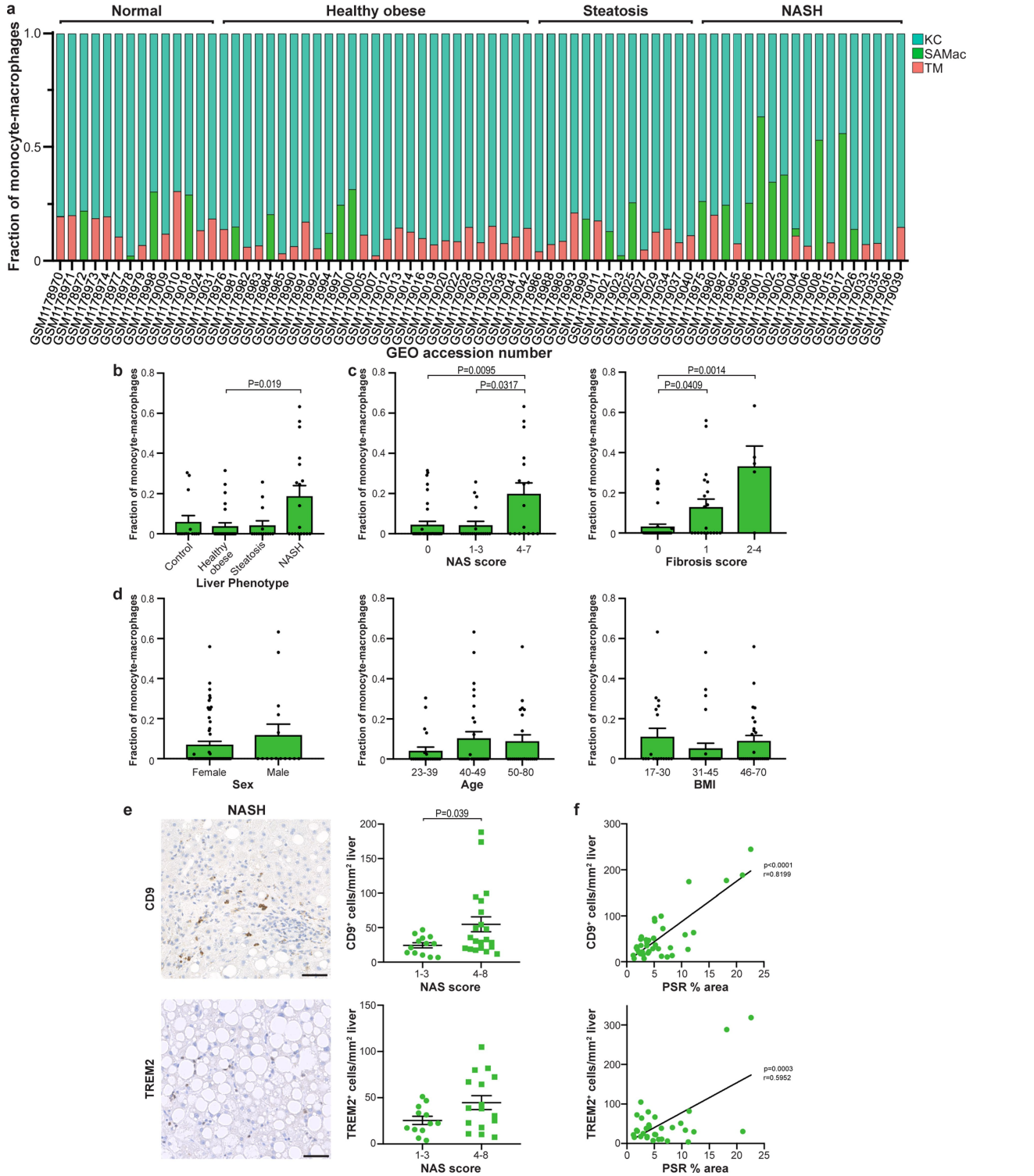


Extended Data Fig. 5 | See next page for caption.

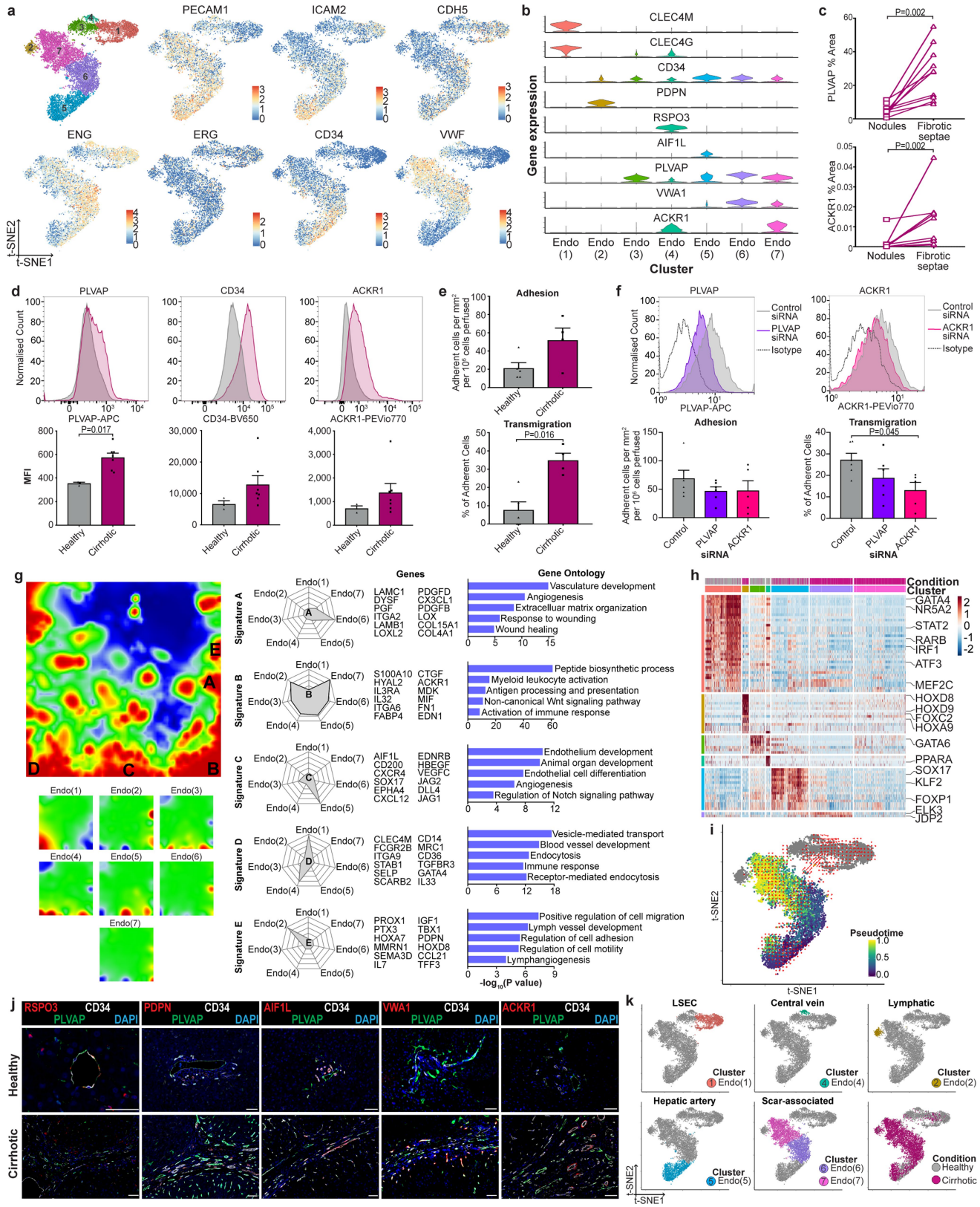
Extended Data Fig. 5 | Phenotypic characterization of mononuclear phagocytes in healthy and cirrhotic human livers. a, Top, self-organizing map (60 × 60 grid) of smoothed scaled metagene expression of 10,737 MPs from healthy ($n = 5$) and cirrhotic ($n = 5$) livers. In total, 20,952 genes, 3,600 metagenes and 44 signatures were identified. A–F denote metagene signatures overexpressed in one or more MP subpopulations. Bottom, smoothed mean metagene expression profile for each MP subpopulation. **b**, Radar plots (left), exemplar genes (middle) and selected GO enrichment (right) of metagene signatures A–F showing distribution of signature expression across MP subpopulations from 10,737 MP cells. **c**, Diffusion map (DM) visualization of blood monocytes and liver-resident MP lineages (23,075 cells from healthy ($n = 5$) and cirrhotic ($n = 5$) liver samples and PBMCs ($n = 5$)), annotating monocle pseudotemporal dynamics (purple to yellow). Top, RNA velocity field (red arrows) visualized using Gaussian smoothing on regular grid. Bottom, annotation of MPs by subpopulation and injury condition. **d**, Unspliced–spliced phase portraits (top); 23,075 cells coloured and visualized as in Fig. 3a; monocyte (*MNDA*), SAMac (*CD9*) and KC (*TIMD4*) marker genes. Cells plotted above or below the steady-state (black dashed line) indicate increasing or decreasing expression of gene, respectively. Spliced expression profile for stated genes (middle row; red, high, blue, low). Unspliced residuals for stated genes (bottom row), positive (red) indicating expected upregulation, negative (blue) indicating expected downregulation. *MNDA* displays negative velocity in SAMacs; *CD9* displays positive velocity in monocytes and SAMacs; *TIMD4* velocity is restricted to KCs. **e**, Cubic smoothing spline curve fitted to averaged expression of all genes in module 2 from the blood monocyte-to-SAMac pseudotemporal trajectory (see Fig. 3c), with selected GO enrichment (right). **f**, Cubic smoothing spline curve fitted to averaged expression of all genes in module 3 from the blood monocyte-to-cDC pseudotemporal trajectory (see Fig. 3c), with selected GO enrichment (right). **g**, Luminex assay showing quantification of levels of stated proteins in culture medium from FACS-isolated SAMacs ($n = 3$), TMs ($n = 2$) and KCs ($n = 2$). Control denotes medium alone ($n = 2$). Data are mean ± s.e.m. **h**, Heat map of transcription factor regulons across MP pseudotemporal trajectory and in KCs (colour-coded by MP cluster, condition and pseudotime), with selected regulons labelled (right). Columns denote cells; rows denote genes. **i**, Scaled regulon expression of selected regulons across MP clusters from healthy ($n = 5$) and cirrhotic ($n = 5$) livers. All *P* values determined by Fisher's exact test.



Extended Data Fig. 6 | Characterization of macrophages in mouse liver fibrosis. a, Clustering and annotating 3,250 mouse (m)MPs from healthy ($n = 3$) and fibrotic (4 weeks CCl_4 treatment; $n = 3$) livers. **b**, Annotating mouse MP cells by injury condition. **c**, Heat map of mouse MP cluster marker genes (top; colour-coded by cluster and condition), with exemplar genes labelled (right). Columns denote cells; rows denote genes. **d**, Selected genes expressed in 3,250 mouse MPs. **e**, Representative flow cytometry plots of the gating strategy ($n = 8$ from two independent experiments) for identifying mouse KCs, $CD9^-$ TMs and $CD9^+$ SAMacs in fibrotic mice. **f**, Quantifying mouse macrophage subpopulations by flow cytometry in healthy ($n = 6$) and fibrotic ($n = 8$) mouse livers from two independent experiments. The macrophage subpopulation (x axis) is shown as a percentage of total viable $CD45^+$ cells (y axis). Data are mean ± s.e.m. P values determined by two-tailed Mann–Whitney test. **g**, Co-culture of primary mouse HSCs from uninjured livers and either FACS-isolated $CD9^-$ mouse TMs or $CD9^+$ mouse SAMacs from fibrotic livers ($n = 8$ mice; two independent experiments). Right, qPCR of *Col3a1* expression in HSCs; expression relative to mean expression of quiescent HSC. P value determined by two-tailed Wilcoxon test. **h**, Clustering 3,250 mouse MPs and 10,737 human (h)MPs into five clusters using canonical correlation analysis. Annotation of cross-species clusters (identity). **i**, Annotation of human and mouse macrophage subpopulations from 3,250 mouse MPs and 10,737 human MPs. **j**, Selected genes expressed in 3,250 mouse MPs and 10,737 human MPs.

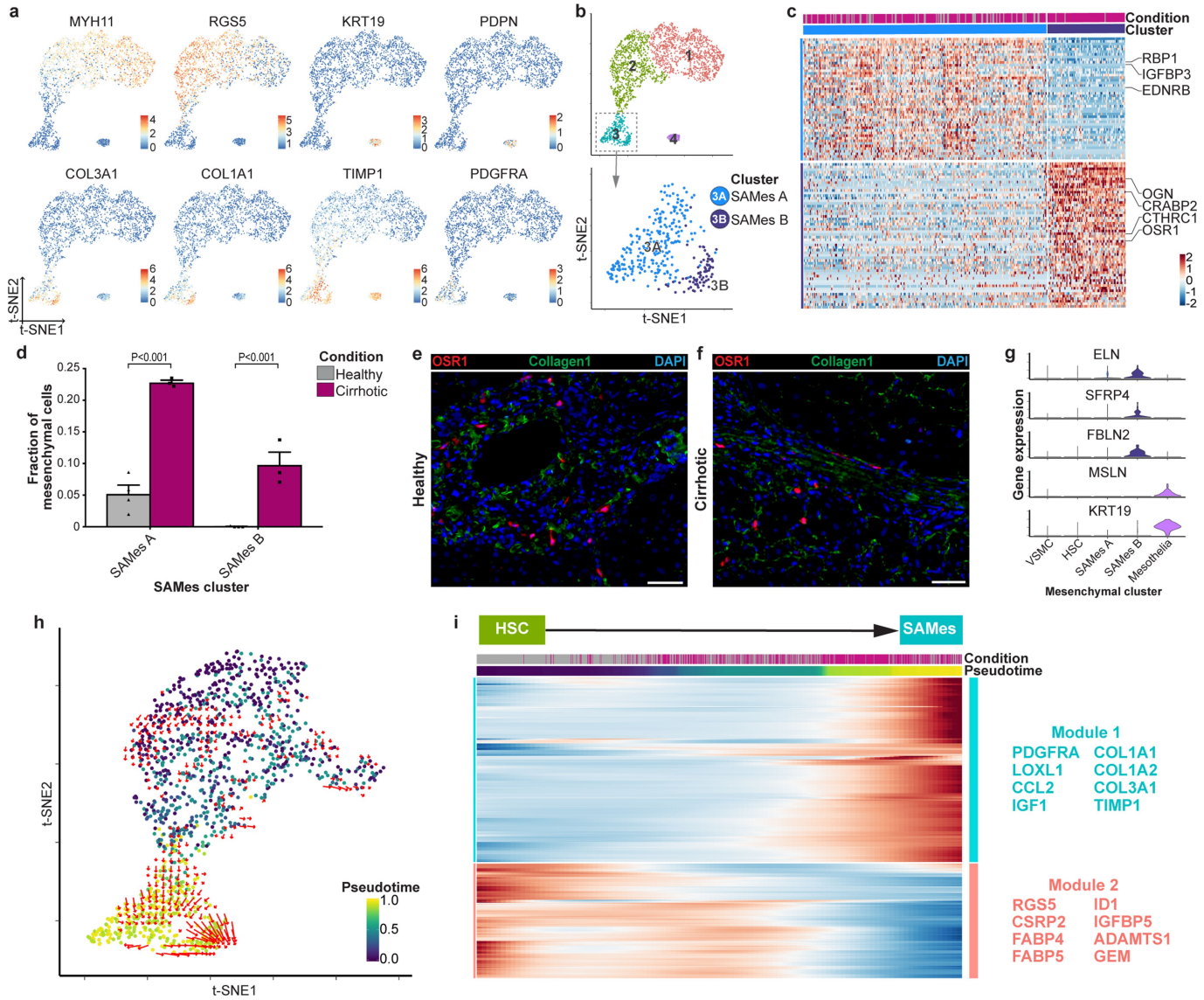


Extended Data Fig. 7 | SAMac expansion in human NASH. a–**d**, Deconvolution of publicly available whole liver microarray data ($n = 73$) assessed for frequency of SAMacs, KCs and TMs using the Cibersort algorithm. **a**, Macrophage composition. GEO accession numbers are shown on the *x* axis; the fraction of monocyte–macrophages is shown on the *y* axis. Liver phenotypes are annotated at the top. **b**, Frequency of SAMacs in control ($n = 14$), heathy obese ($n = 27$), steatosis ($n = 14$) and NASH ($n = 18$) livers. **c**, Left, frequency of SAMacs in patients with histological NAFLD activity scores (NAS) of 0 ($n = 37$), 1–3 ($n = 19$) and 4–7 ($n = 17$). Right, frequency of SAMacs in patients with histological fibrosis scores of 0 ($n = 46$), 1 ($n = 20$) and 2–4 ($n = 5$). **d**, Left, frequency of SAMacs in female ($n = 58$) and male ($n = 15$) patients. Middle, frequency of SAMacs in patients aged 23–39 ($n = 22$), 40–49 ($n = 29$) and 50–80 ($n = 22$). Right, frequency of SAMacs in patients with a body mass index (BMI) of 17–30 ($n = 18$), 31–45 ($n = 28$) and 46–70 ($n = 27$). **e**, Left, immunohistochemistry of CD9 and TREM2 expression in NAFLD liver biopsy sections. Scale bars, 50 µm. Right, cell counts of CD9 and TREM2 expression. CD9: NAS 1–3 ($n = 13$), NAS 4–8 ($n = 21$). TREM2: NAS 1–3 ($n = 12$), NAS 4–8 ($n = 16$). **f**, Correlation of cell counts with picrosirius red (PSR) digital morphometric pixel quantification in NAFLD liver biopsy tissue with CD9 staining (top; $n = 39$) or TREM2 staining (bottom; $n = 32$). Data are mean ± s.e.m. *P* values determined by Kruskal–Wallis and Dunn test (**b**, **c**), two-tailed Mann–Whitney test (**e**) or Pearson's correlation and linear regression (**f**).

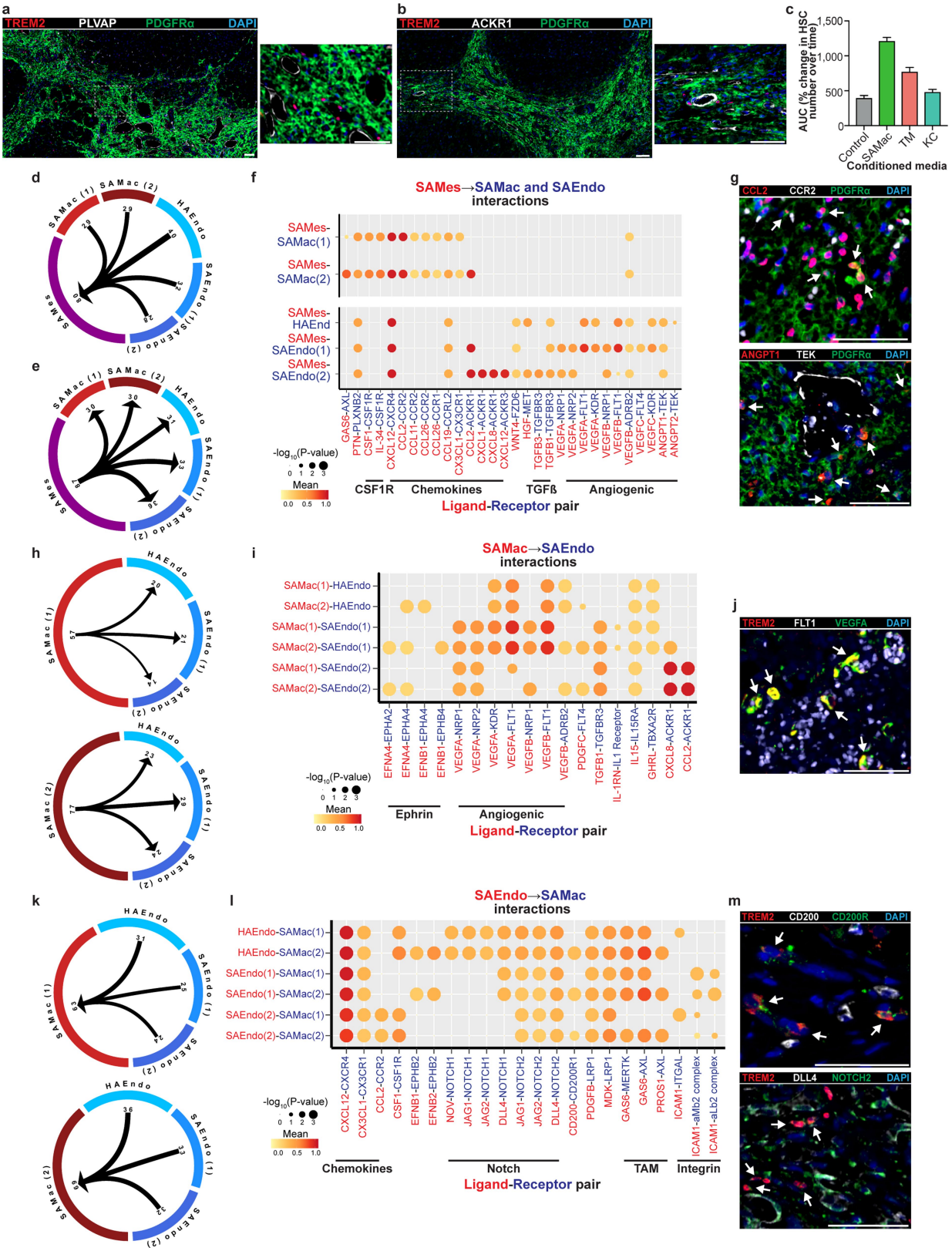


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Phenotypic characterization of endothelial cells in healthy and cirrhotic human livers. a, Clustering and selected genes expressed in 8,020 endothelial cells from healthy ($n = 4$) and cirrhotic ($n = 3$) human livers. **b**, Scaled gene expression of endothelial cluster markers across endothelial cells from healthy ($n = 4$) and cirrhotic ($n = 3$) livers. **c**, Top, digital pixel quantification of PLVAP immunostaining of cirrhotic liver sections ($n = 10$) in parenchymal nodules and fibrotic septae. Bottom, ACKR1 immunostaining of cirrhotic liver sections ($n = 10$) in parenchymal nodules and fibrotic septae. **d**, Flow cytometry analysis of PLVAP, CD34 and ACKR1 in endothelial cells from healthy ($n = 3$, grey) or cirrhotic ($n = 7$, red) livers. Top, representative histograms; bottom, MFI values. **e**, Flow-based adhesion assay. Peripheral blood monocytes assessed for adhesion to primary human liver endothelial cells (top) and the percentage of adherent monocytes that transmigrate (bottom); endothelial cells isolated from healthy ($n = 5$) or cirrhotic ($n = 4$) livers. **f**, Endothelial cell gene knockdown. Cirrhotic endothelial cells were treated with siRNA against *PLVAP* ($n = 6$) or *ACKR1* ($n = 5$) or with control siRNA ($n = 6$). Top, representative flow cytometry histograms for stated markers, with comparison to isotype control antibody. Bottom, flow-based adhesion assay, with PBMCs assessed for adhesion (bottom left) and the percentage of adherent cells that transmigrate (bottom right) after siRNA treatment of endothelial cells. **g**, Top left, self-organizing map (60 × 60 grid) of smoothed scaled metagene expression of endothelia lineage. In total, 21,237 genes, 3,600 metagenes and 45 signatures were identified. A–E denote metagene signatures overexpressed in one or more endothelial subpopulations. Bottom left, smoothed mean metagene expression profile for each endothelial subpopulation. Middle, radar plots of metagene signatures A–E showing distribution of signature expression across endothelial subpopulations, exemplar genes (middle) and Gene Ontology enrichment (right). **h**, Heat map of endothelial subpopulation transcription factor regulon expression (colour-coded by cluster and condition) across 8,020 endothelial cells from healthy ($n = 4$) and cirrhotic ($n = 3$) human livers. Exemplar regulons are labelled (right). Columns denote cells; rows denote regulons. **i**, *t*-SNE visualization of endothelial lineage (8,020 cells from healthy ($n = 4$) and cirrhotic ($n = 3$) livers), annotating monocle pseudotemporal dynamics (purple to yellow; grey indicates lack of inferred trajectory). RNA velocities (red arrows) visualized using Gaussian smoothing on regular grid. **j**, Representative immunofluorescence images ($n \geq 3$) of RSPO3, PDPN, AIF1L, VWA1 or ACKR1 (red), CD34 (white), PLVAP (green) and DAPI (blue) in healthy and cirrhotic liver. Scale bars, 50 µm. **k**, Annotation of 8,020 endothelial cells by subpopulation and injury condition. LSEC, liver sinusoidal endothelial cells. Data are mean ± s.e.m. *P* values determined by two-tailed Wilcoxon test (**c**), two-tailed Mann–Whitney test (**d**, **e**), Kruskal–Wallis and Dunn test (**f**), or Fisher's exact test (**g**).



Extended Data Fig. 9 | Characterization of mesenchymal cells in healthy and cirrhotic human livers. a, Selected genes expressed in 2,318 mesenchymal cells from healthy ($n = 4$) and cirrhotic ($n = 3$) human livers. **b**, Clustering 319 SAMes into two further subclusters. **c**, Heat map of SAMes subcluster marker genes (colour-coded by cluster and condition), with exemplar genes labelled (right). Columns denote cells; rows denote genes. **d**, Fractions of SAMes subpopulations in healthy ($n = 4$) and cirrhotic ($n = 3$) livers. **e**, **f**, Representative immunofluorescence images ($n \geq 3$) of OSR1 (red), collagen 1 (green) and DAPI (blue) in portal region of healthy liver (**e**) or fibrotic niche of cirrhotic liver (**f**). Scale bars, 50 µm. **g**, Scaled gene expression of selected genes across 2,318 mesenchymal cells from healthy ($n = 4$) and cirrhotic ($n = 3$) livers. **h**, *t*-SNE visualization of 1,178 HSCs and SAMes from healthy ($n = 4$) and cirrhotic ($n = 3$) livers annotated by monocle pseudotemporal dynamics (purple to yellow). RNA velocity field (red arrows) visualized using Gaussian smoothing on regular grid. **i**, Heat map of cubic smoothing spline curves fitted to genes differentially expressed across HSC-to-SAMes pseudotemporal trajectories, grouped by hierarchical clustering ($k = 2$); colour-coded by pseudotime and condition (top). Gene co-expression modules (colour) and exemplar genes are labelled (right). Data are mean ± s.e.m. *P* values determined by Wald test (**d**).



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Characterization of the cellular interactome in the fibrotic niche. a, **b**, Representative immunofluorescence images ($n \geq 3$) of fibrotic niche in cirrhotic liver. **a**, TREM2 (red), PLVAP (white), PDGFRα (green) and DAPI (blue). **b**, TREM2 (red), ACKR1 (white), PDGFRα (green) and DAPI (blue). **c**, Proliferation assay. Human HSCs were treated with conditioned medium from primary hepatic macrophage subpopulations SAMac ($n = 2$), TMs ($n = 2$), KCs ($n = 2$) or control medium ($n = 2$). The AUC of the percentage change in HSC number over time (hours) is shown on the *y* axis. Data are mean ± s.e.m. **d**, Circle plot showing potential interaction magnitude from ligands expressed by SAMacs and SAEndos to receptors expressed on SAMes. **e**, Circle plot showing potential interaction magnitude from ligands expressed by SAMes to receptors expressed on SAMacs and SAEndos. **f**, Dot plot of ligand–receptor interactions between SAMes ($n = 7$ human livers), SAMacs ($n = 10$ human livers) and SAEndos ($n = 7$ human livers). Ligand (red) and cognate receptor (blue) shown on the *x* axis; populations that express ligand (red) and receptor (blue) are shown on the *y* axis; circle size denotes *P* value (permutation test); colour (red, high; yellow, low) denotes average ligand and receptor expression levels in interacting subpopulations. **g**, Top, representative immunofluorescence image ($n \geq 3$) of CCL2 (red), CCR2 (white), PDGFRα (green) and DAPI (blue) in fibrotic niche in cirrhotic liver; arrows denote $CCL2^+PDGFR\alpha^+$ cells. Bottom, representative immunofluorescence image ($n \geq 3$) of ANGPT1 (red), TEK (white), PDGFRα (green) and DAPI (blue) in fibrotic niche in cirrhotic liver; arrows denote $ANGPT1^+PDGFR\alpha^+$ cells. **h**, Circle plot denotes potential interaction magnitude from ligands expressed by SAMacs to receptors expressed on SAEndos. **i**, Dot plot of ligand–receptor interactions between SAMacs ($n = 10$ human livers) and SAEndos ($n = 7$ human livers) as in **f**. **j**, Representative immunofluorescence image ($n \geq 3$) of TREM2 (red), FLT1 (white), VEGFA (green) and DAPI (blue) in fibrotic niche in cirrhotic liver; arrows denote $TREM2^+VEGFA^+$ cells. **k**, Circle plot of the potential interaction magnitude from ligands expressed by SAEndos to receptors expressed on SAMacs. **l**, Dot plot of ligand–receptor interactions between SAEndo ($n = 7$ human livers) and SAMacs ($n = 10$ human livers) as in **f**. **m**, Top, representative immunofluorescence image ($n \geq 3$) of TREM2 (red), CD200 (white), CD200R (green) and DAPI (blue) in fibrotic niche in cirrhotic liver; arrows denote $TREM2^+CD200R^+$ cells. Bottom, representative immunofluorescence image ($n \geq 3$) of TREM2 (red), DLL4 (white), NOTCH2 (green) and DAPI (blue) in fibrotic niche in cirrhotic liver; arrows denote $TREM2^+NOTCH2^+$ cells. All scale bars, 50 µm.

nature research

Corresponding author(s): Prakash Ramachandran
Neil C. Henderson

Last updated by author(s): Aug 20, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Initial processing of single-cell RNA-sequencing data was performed using the commercial CellRanger pipeline (10X Genomics, version 2.1.0, see Methods). Subsequent analysis was performed using the open-source R programming language (version 3.4.1). BD FACS Sortware software was used for cell sorting on BD Influx equipment. BD FACS Diva software was used for flow cytometry on BD LSR Fortessa equipment and for cell sorting on BD FACSAria Fusion and FACSAriaII. Fluorescent and brightfield microscopy images were acquired using Zen Blue software (Zeiss) on an Axioscan.Z1 instrument (Zeiss) or Confocal Microscope Zeiss LSM780. Luminex data was acquired on a Bio-Plex 200 (Bio-Rad). Cell Proliferation data was acquired on an Incucyte ZOOM live cell analysis system (Essen biosciences). RT-qPCR data was acquired on ABI 7900HT FAST PCR system (Applied Biosystems).
Data analysis	Immunoflurorescent images were processed and analysed using Zen Blue software (Zeiss) and Fiji software (ImageJ version 2.00). Cell proliferation data were analysed on the Incucyte proprietary analysis software (version 2018A). Immunohistochemistry images were analysed using QuPAth software (version 0.1.2) for automated cell counting and using Fiji software (ImageJ version 2.00) with Trainable Weka Segmentation plugin (see Methods). Co-culture immunocytochemistry data was analysed using Imaris x64 (version 8.1.2). Flow cytometry analysis was performed using FlowJo software (version 10.2). RT-qPCR data was analysed using ThermoFisher Connect cloud qPCR software (version 2019.1.8-Q1-19-build4). Statistical analysis was performed either in R (version 3.4.1) or using Graphpad Prism software (version 7.0a). Single-cell RNA-sequencing analysis was performed in R, based around the following packages: Seurat 2.3.0, scImpute 0.0.8, SCRAT 1.0.0, monocle 2.6.1, scater 1.4.0, velocyto 0.6.0, SCENIC 0.1.7 (see Methods). We also made use of the CellPhoneDB repository of ligands, receptors, and interactions. Deconvolution was performed using Cibersort. Gene Ontology enrichment analysis was performed using PANTHER 13.1.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Our expression data will be freely available for user-friendly interactive browsing online at www.livercellatlas.mvm.ed.ac.uk. CellPhoneDB is available at www.CellPhoneDB.org, along with lists of membrane proteins, ligands and receptors, and heteromeric complexes. All raw sequencing data have been deposited in the Gene Expression Omnibus (GEO Accession GSE136103). We make available as Supplementary Tables: lists of lineage-specific genes for signature analysis (Extended Data Figure 1e, 2b); lists of marker genes and regulons from clustering results (Figure 1e, 2d, 4c, 5b, Extended Data Figure 3d, e, g, 5h, 6c, 8h, 9c); lists of module / signature genes from trajectory and self-organising map analyses and corresponding lists of gene ontology terms from enrichment analysis (Figure 3c, d, Extended Data Figure 5a, b, e, f, 8g); lists of significant interactions in the fibrotic niche as identified using CellPhoneDB (Figure 6a, e, Extended Data Figure 10f, i, l).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	In total, we present scRNA-seq data from ten human liver samples (n=5 healthy and n=5 cirrhotic), five human blood samples (n=4 cirrhotic and n=1 healthy named PBMC8K; pbmc8k dataset sourced from single-cell gene expression datasets hosted by 10X Genomics) and two mouse samples (n=3 healthy and n=3 fibrotic). No statistical methods were used to predetermine sample size. Patient number was selected to give a balanced representation of healthy and cirrhotic liver cells and to provide sufficient cells of each lineage to facilitate more detailed analysis. Histology, flow cytometry, luminex, RT-qPCR and cell proliferation analysis were performed on multiple independent biological replicates (n shown in figure legends).
Data exclusions	Described in detail in Methods. Exclusion criteria were determined following initial assessment and QC of the data. Low gene expression (fewer than 300 genes) or a high mitochondrial gene content (>30% of the total UMI count) are indicators of outlier low quality cells and were excluded. At each stage of the analysis we used signature analysis to identify and exclude potential doublet clusters.
Replication	All experimental findings reported here were successfully replicated across multiple biological samples (n reported in each figure legend). All immunofluorescence was performed on a minimum of 3 liver samples to identify representative images.
Randomization	One group of randomly selected healthy livers and another group of randomly selected cirrhotic livers were analysed in this study. All subsequent analyses were performed in randomly selected healthy or cirrhotic liver samples. For mouse experiments, age-matched littermate mice were randomly assigned to be healthy controls or receive carbon tetrachloride.
Blinding	Blinding to the origin of the tissue samples was not possible. All analyses were performed in an automated manner across conditions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	All antibodies used in this work, clone, application, supplier and lot number are listed in Supplementary Table 19.
Validation	All antibodies used are commerically available and validated by the vendor for the assay and species used in this study. Specific validation information for each antibody is available on the vendors website. The specificity of each primary flow cytometry antibody was validated by staining directly against species-matched isotype and unstained controls. Validation of each primary antibody used for immunostaining was performed by comparison to species-matched isotype antibodies and unstained controls

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Male C57BL/6JCrl mice aged 8 to 10 weeks
Wild animals	Study did not involve wild animals
Field-collected samples	Study did not involve samples collected in the field
Ethics oversight	All experiments were performed in accordance with UK Home Office regulations.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Please see Extended Data Figure 1a for the clinical characteristics of patients used for single-cell RNA sequencing.
Recruitment	Patients were recruited as described in Methods. Healthy background non-lesional liver tissue was obtained intraoperatively from patients undergoing surgical liver resection for solitary colorectal metastasis at the Hepatobiliary and Pancreatic Unit, Department of Clinical Surgery, Royal Infirmary of Edinburgh. Patients with a known history of chronic liver disease, abnormal liver function tests or those who had received systemic chemotherapy within the last four months were excluded from this cohort. Cirrhotic liver tissue was obtained intraoperatively from patients undergoing orthotopic liver transplantation at the Scottish Liver Transplant Unit, Royal Infirmary of Edinburgh. Blood from patients with a confirmed diagnosis of liver cirrhosis were obtained from patients attending the Scottish Liver Transplant Unit, Royal Infirmary of Edinburgh. Patients with liver cirrhosis due to viral hepatitis were excluded from the study. For cell sorting of macrophages or isolation of human endothelial cells, liver tissue was acquired from explanted diseased livers from patients undergoing orthotopic liver transplantation, resected liver specimens or donor livers rejected for transplant at the Queen Elizabeth Hospital, Birmingham.
Ethics oversight	NRS BioResource and Tissue Governance Unit (Study Number SR574), following review at the East of Scotland Research Ethics Service (Reference 15/ES/0094) For University of Birmingham samples, separate local ethical approval was obtained (Reference 06/Q2708/11, 06/Q2702/61).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Please see Methods for detailed sample preparation protocol for FACS and flow cytometry
Instrument	BD Influx and BD FACSAriaII were used for cell sorting at University of Edinburgh. BD LSR Fortessa was used for flow cytometry analysis. BD FACSAria Fusion for cell sorting at the University of Birmingham
Software	BD FACS Sortware software was used for cell sorting on BD Influx equipment. BD FACS Diva software for cell sorting on BD FACSAriaII and BD FACSAria Fusion. BD FACS Diva software was used for flow cytometry on BD LSR Fortessa equipment. Flow

cytometry analysis was performed using FlowJo software (version 10.2).

Cell population abundance

Sort purity was routinely over 95% on post-sort checks

Gating strategy

Please see Methods and Extended Data Figures 1b,c, 4f and 6e for gating strategies. Initial gating for all experiments: Cells (FSC-A vs SSC-A), Singlets (FSC-A vs FSC-H (or FSC-A vs TPW for BD Influx)), Viable (SSC-A vs viability dye (See methods)). For human PBMC sort, CD45+ CD66b- cells were sorted. For human liver single-cell RNA-seq sorting, CD45+ cells (leukocytes) or CD45- cells (other NPCs) were sorted. For human liver macrophage flow cytometry quantification and cell sorting, tissue monocyte-macrophages were identified as CD45+, Lin- (CD3, CD335, CD19, CD66b, LILRA4, CD326), HLA-DR+, CD1C-, CD14+ and/or CD16+ cells. SAM were then identified as CD163- TREM2+ CD9+, KCs were identified as CD163+ CD9- and TMo were identified as CD163-. For mouse liver single-cell RNA-seq sorting, tissue mononuclear phagocytes identified as CD45+ Lin-(CD3, NK1.1, Ly6G, CD19) cells were sorted. For mouse liver macrophage cell sorting, CD45+ Lin- CD11b+ F4/80+ TIMD4- CD9+ or CD9- cells were sorted from CCl4-treated mice. For human liver endothelial cell flow cytometry, cultured endothelial cells were stained with antibodies to PLVAP, ACKR1, JAG1 and CD34. Gates and boundaries were defined by comparison to FMO and unstained samples.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Article

Immunity to commensal papillomaviruses protects against skin cancer

https://doi.org/10.1038/s41586-019-1719-9

Received: 20 November 2018

Accepted: 3 September 2019

Published online: 30 October 2019

John D. Strickley[1,2,3,7], Jonathan L. Messerschmidt[1,7], Mary E. Awad[1], Tiancheng Li[1], Tatsuya Hasegawa[1], Dat Thinh Ha[1,2,3], Henry W. Nabeta[2,3], Paul A. Bevins[1,2,3], Kenneth H. Ngo[1], Maryam M. Asgari[4], Rosalynn M. Nazarian[5], Victor A. Neel[4], Alfred Bennett Jenson[2], Joongho Joh[2,3,6] & Shadmehr Demehri[1,4]*

Immunosuppression increases the risk of cancers that are associated with viral infection[1]. In particular, the risk of squamous cell carcinoma of the skin—which has been associated with beta human papillomavirus (β-HPV) infection—is increased by more than 100-fold in immunosuppressed patients[2–4]. Previous studies have not established a causative role for HPVs in driving the development of skin cancer. Here we show that T cell immunity against commensal papillomaviruses suppresses skin cancer in immunocompetent hosts, and the loss of this immunity—rather than the oncogenic effect of HPVs—causes the markedly increased risk of skin cancer in immunosuppressed patients. To investigate the effects of papillomavirus on carcinogen-driven skin cancer, we colonized several strains of immunocompetent mice with mouse papillomavirus type 1 (MmuPV1)[5]. Mice with natural immunity against MmuPV1 after colonization and acquired immunity through the transfer of T cells from immune mice or by MmuPV1 vaccination were protected against skin carcinogenesis induced by chemicals or by ultraviolet radiation in a manner dependent on $CD8^+$ T cells. RNA and DNA in situ hybridization probes for 25 commensal β-HPVs revealed a significant reduction in viral activity and load in human skin cancer compared with the adjacent healthy skin, suggesting a strong immune selection against virus-positive malignant cells. Consistently, E7 peptides from β-HPVs activated $CD8^+$ T cells from unaffected human skin. Our findings reveal a beneficial role for commensal viruses and establish a foundation for immune-based approaches that could block the development of skin cancer by boosting immunity against the commensal HPVs present in all of our skin.

Cutaneous squamous cell carcinoma (SCC) is the second-most-common type of cancer and is associated with severe morbidity and mortality—especially among immunosuppressed patients such as recipients of organ transplants[2]. Although ultraviolet (UV) radiation is the main and preventable cause of skin cancer, the incidence of skin cancer in the United States has doubled from 1992 to 2012[2], highlighting the urgent need to develop new approaches for the prevention and treatment of this disease. Given that β-HPVs have been found in more than 80% of SCCs among recipients of organ transplants, a potential viral cause of skin cancer has been proposed[2–4]. However, unlike high-risk α-HPVs, no predominant types of β-HPV have been identified in skin cancers, and the β-HPV genome is rarely integrated into the DNA of cancer cells and is not transcriptionally active[6]. Findings like these have led to a 'hit-and-run' hypothesis, in which β-HPV facilitates the initiation of UV-driven skin cancer but is later lost during tumour maintenance[4,7].

To investigate the role of papillomavirus in carcinogen-driven skin cancer, we used a MmuPV1 back-skin infection system, which led to the development of confluent warts in $Cd4^{-/-}Cd8^{-/-}$ (*Cd8* is also known as *Cd8a*) mice but no skin lesions in immunocompetent, wild-type mice (Extended Data Fig. 1a–c). Two months after infection, MmuPV1- and sham-infected mice of the C57BL/6J strain were subjected to a chemical carcinogenesis protocol for 30 weeks. Notably, MmuPV1-colonized mice showed a significant delay in the onset of skin tumours, developed significantly fewer tumours over time and completed the study with a significantly lower tumour burden, compared with sham-infected controls (Fig. 1a–c, Extended Data Fig. 1d). In the FVB strain, 23% of MmuPV1-infected wild-type mice showed complete immunity five weeks after infection (that is, no skin warts; Fig. 1d, Extended Data Fig. 1e). At ten weeks after infection, warts had completely regressed in 58% of the wart-bearing mice—indicative of antiviral adaptive immunity. Adoptive transfer of memory T cells from mice that are immune to MmuPV1 into $Cd4^{-/-}Cd8^{-/-}$ mice led to fewer warts in these mice after MmuPV1 back-skin infection than were observed in control T cell-deficient mice and $Cd4^{-/-}Cd8^{-/-}$ mice that received memory T cells from wild-type mice

[1]Center for Cancer Immunology and Cutaneous Biology Research Center, Center for Cancer Research, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [2]James Graham Brown Cancer Center, University of Louisville, Louisville, KY, USA. [3]Department of Medicine, University of Louisville, Louisville, KY, USA. [4]Department of Dermatology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [5]Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [6]Center for Predictive Medicine, University of Louisville, Louisville, KY, USA. [7]These authors contributed equally: John D. Strickley, Jonathan L. Messerschmidt. *e-mail: sdemehri1@mgh.harvard.edu

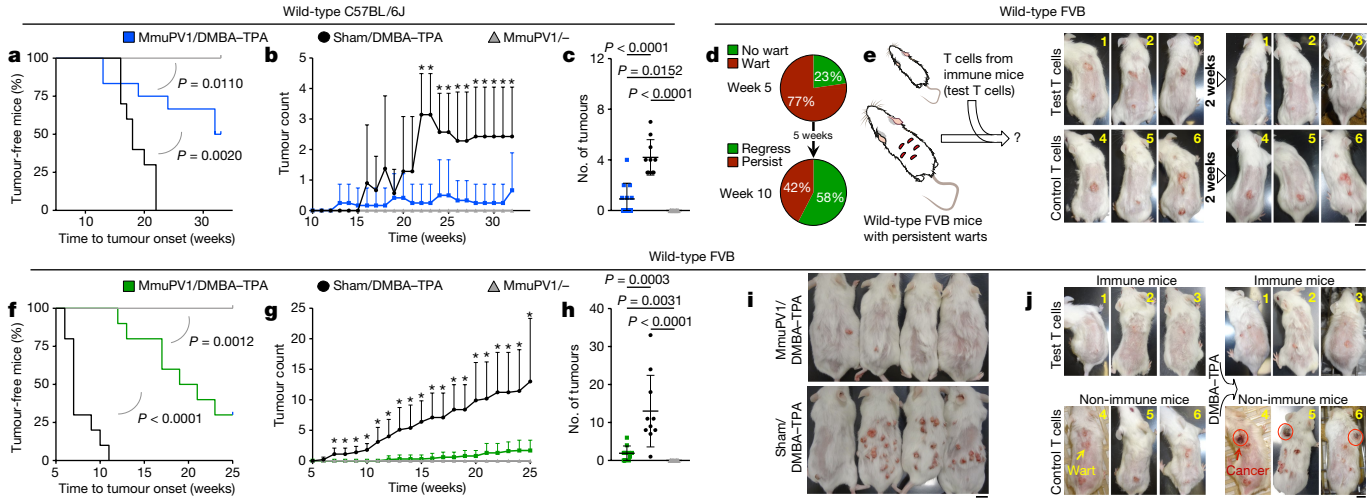


Fig. 1 | Colonization of the skin with MmuPV1 protects mice against chemically induced skin carcinogenesis. a–**c**, Skin tumour outcomes for wild-type C57BL/6J mice infected with MmuPV1 and treated with DMBA–TPA (MmuPV1/DMBA–TPA) ($n = 12$); sham-infected and treated with DMBA–TPA (sham/DMBA–TPA) ($n = 10$); or infected with MmuPV1 alone (MmuPV1/–) ($n = 10$). **a**, Time to tumour onset (log-rank test). **b**, Tumour count per mouse over time. **c**, Number of tumours per mouse at the completion of the carcinogenesis protocol. Two-tailed Mann–Whitney *U* test (**b**, **c**); $^*P < 0.05$. **d**, Percentage of wild-type FVB mice with wart development on their back skin at five weeks, and a subgroup with persistent warts at ten weeks after infection with MmuPV1. **e**, T cells from skin-draining lymph nodes of MmuPV1-colonized immune mice (test T cells) transferred to mice with persistent warts ($n = 3$). The change in the burden of skin warts is shown at two weeks after adoptive transfer of T cells. In the control group ($n = 3$), T cells represent naive T cells from the spleen of uninfected wild-type FVB mice. **f**–**i**, Development of skin tumours in wild-type FVB MmuPV1/DMBA–TPA mice ($n = 10$) compared with sham/DMBA–TPA ($n = 10$) and MmuPV1/– ($n = 10$) mice. **f**, Time to tumour onset (log-rank test). **g**, Tumour count per mouse over time. **h**, Number of tumours per mouse at the completion of the carcinogenesis protocol. Two-tailed Mann–Whitney *U* test (**g**, **h**); $^*P < 0.05$. **i**, Representative images of mice in DMBA–TPA-treated cohorts. **j**, Mice that rejected their skin warts after receiving T cells from MmuPV1-colonized immune mice (test T cells in **e**; $n = 3$) compared with wart-bearing non-immune mice ($n = 3$) after treatment with DMBA–TPA. Note that non-immune mice developed invasive skin cancers at 8 weeks (mouse 4), 15 weeks (mouse 5) and 20 weeks (mouse 6) after DMBA–TPA treatment (red circles). Mice were shaved to enable better visualization of the skin tumours. Data represent two independent sets of experiments with similar results. Data are mean + s.d. (**b**, **g**) or mean ± s.d. (**c**, **h**). Scale bars, 1 cm (**e**, **i**, **j**).

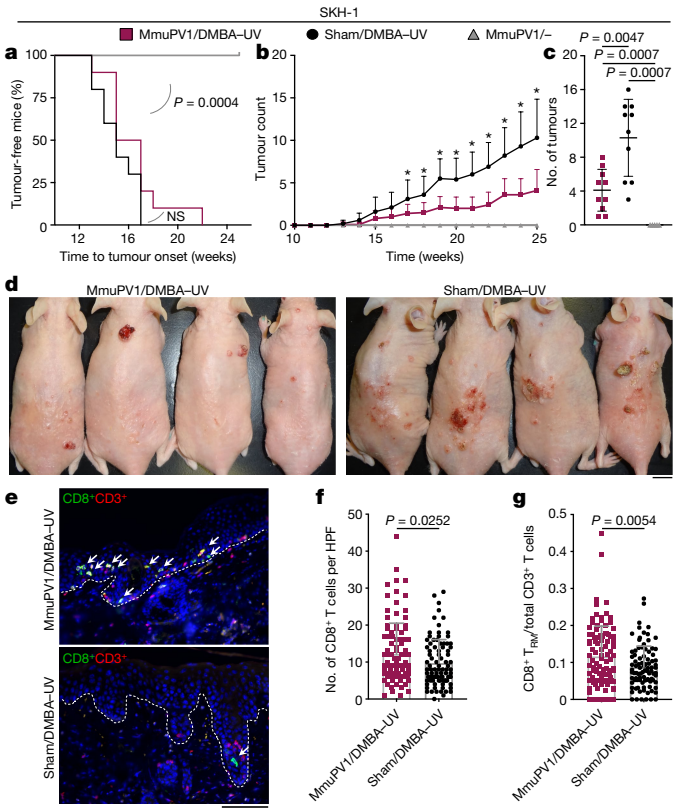


Fig. 2 | Colonization of the skin with MmuPV1 protects immunocompetent SKH-1 mice against UV-induced carcinogenesis. a–**c**, SKH-1 mice with MmuPV1-colonized back skin and no warts (that is, immune) were subjected to the DMBA–UV carcinogenesis protocol (Methods). **a**, Time to tumour onset (log-rank test; NS, not significant). **b**, Tumour count per mouse over time. **c**, Number of tumours per mouse at the completion of the carcinogenesis protocol, compared between the following groups: SKH-1 mice infected with MmuPV1 and treated with DMBA–UV (MmuPV1/DMBA–UV) ($n = 10$); sham-infected and treated with DMBA–UV (sham/DMBA–UV) ($n = 10$); or infected with MmuPV1 alone (MmuPV1/–) ($n = 5$). Two-tailed Mann–Whitney *U* test (**b**, **c**); $^*P < 0.05$. **d**, Representative images of mice in DMBA–UV-treated cohorts. Note the resemblance of DMBA–UV-induced skin tumours to actinic keratosis and SCC in humans. Scale bar, 1 cm. **e**, Representative images of $CD8^+$ T cells in the skin of MmuPV1/DMBA–UV and sham/DMBA–UV mice at the completion of the carcinogenesis protocol. Arrows indicate $CD8^+$ T cells in the epidermis; dashed lines highlight the epidermal basement membrane. Scale bar, 100 μm. **f**, $CD8^+$ T cell infiltrates in MmuPV1/DMBA–UV ($n = 10$) and sham/DMBA–UV ($n = 9$) skin, quantified in ten randomly selected high-power field (HPF) images per mouse and averaged across the mice in each group (two-tailed unpaired *t*-test). **g**, The ratio of epidermal $CD8^+$ T cells (that is, $CD8^+ T_{RM}$ cells) to total T cells in each HPF image calculated across MmuPV1/DMBA–UV ($n = 10$) and sham/DMBA–UV ($n = 9$) groups (two-tailed unpaired *t*-test). Each dot represents one high-power image. Stained cells were counted blindly. Data represent two independent sets of experiments with similar results. Data are mean + s.d. (**b**, **f**, **g**) or mean ± s.d. (**c**).

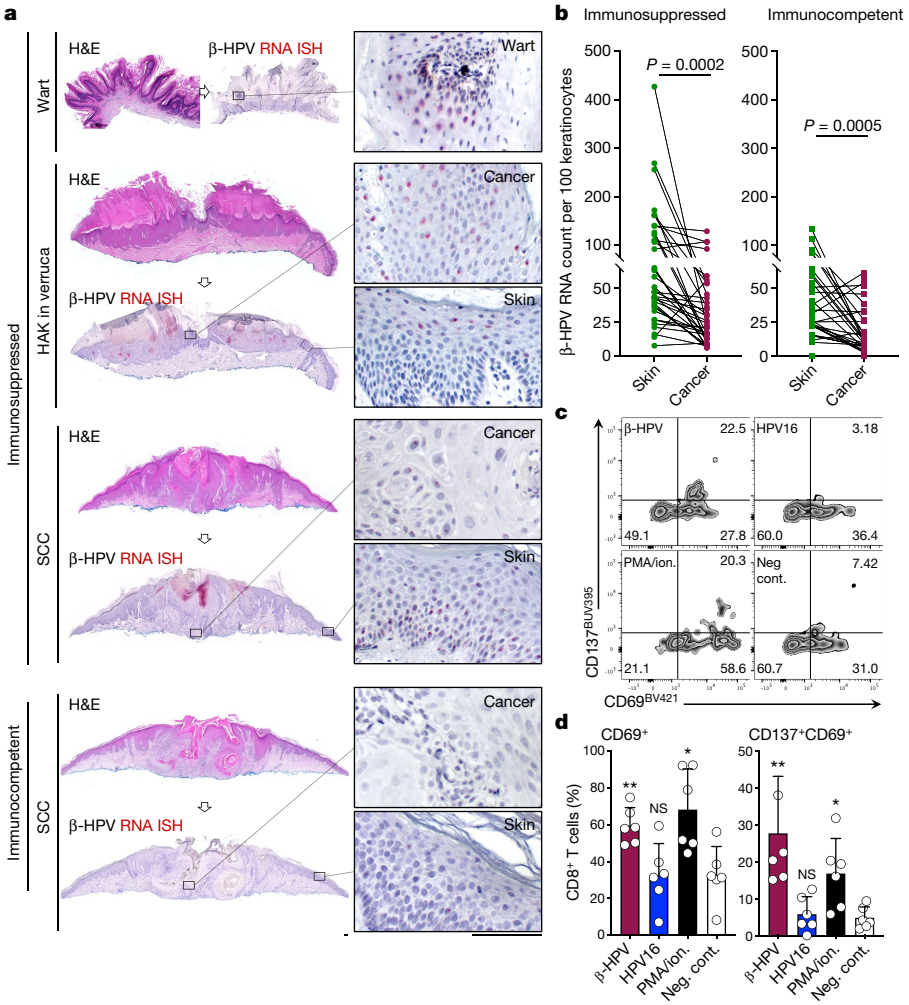


Fig. 3 | Reduced β-HPV transcripts in skin cancer cells and presence of β-HPV-specific CD8⁺ T cells in healthy human skin indicates a selective pressure by antiviral immunity against malignant cells with active HPV. **a**, Representative SCC sections from immunosuppressed and immunocompetent patients, stained with haematoxylin and eosin (H&E) or by β-HPV RNA ISH (red dots). The wart sample serves as a positive control and exhibits the greatest amount of β-HPV activity. Hypertrophic actinic keratosis (HAK) arising in association with a wart (HAK in verruca) is another example of a β-HPV-active lesion found on the skin of immunosuppressed patients. Insets highlight the representative areas of the cancer or wart and their adjacent normal skin. Scale bars, 100 µm. **b**, β-HPV RNA ISH signals quantified in paired samples of skin cancer and its adjacent normal skin, collected from immunosuppressed ($n = 38$) and immunocompetent ($n = 32$) patients (two-tailed paired t-test). Skin cancer characteristics are listed in Supplementary Table 2. **c**, **d**, Representative flow cytometry plots (**c**) and quantification (**d**) of $CD69^+CD8^+$ and $CD137^+CD69^+CD8^+$ T cells isolated from human facial skin and used in a β-HPV peptide stimulation assay ($n = 6$ biological replicates for each treatment condition). T cells were isolated from 8 samples of facial skin (6 males and 2 females; average age 75; age range 60–89). The percentage of $CD8^+$ T cells in each quadrant is shown on the flow cytometry plots. Stimulation with phorbol-12-myristate-13-acetate and ionomycin (PMA/ion.) was used as a positive control. For details of the peptide pool, see Methods and Supplementary Table 3. Two-tailed Mann–Whitney U test; $^*P < 0.05$, $^{**}P < 0.01$, NS, not significant compared with negative control. RNA ISH signals were counted blindly. Data are mean + s.d. (**d**).

treated with parvovirus vaccine (Extended Data Fig. 2a, b). T cells from MmuPV1-immune mice also provided immunity (that is, wart rejection) to wild-type FVB mice with persistent warts (Fig. 1e). MmuPV1-colonized immune FVB mice that received 7,12-dimethylbenz[*a*]anthracene (DMBA) and 12-*O*-tetradecanoylphorbol-13-acetate (TPA) for 20 weeks were protected against chemical carcinogenesis compared with sham-infected mice (Fig. 1f–i). Furthermore, mice with acquired immunity after T cell transfer were also protected from chemical carcinogenesis (Fig. 1j). The MmuPV1 specificity of the transferred T cells from MmuPV1-immune mice was further substantiated by their inability to protect against the growth of SCC cells that were not infected with MmuPV1 (Extended Data Fig. 2c). At the completion of the carcinogenesis studies, MmuPV1 viral DNA and anti-MmuPV1 antibodies were detectable in the normal skin and blood, respectively, of MmuPV1-colonized mice (Extended Data Fig. 3a–d). Although there was no change in overall levels of immune cell infiltrates, MmuPV1-colonized skin had an increased ratio of $CD8^+$ tissue-resident memory T (T_{RM}) cells in the epidermis to total T cells in the skin (Extended Data Fig. 3e–h). DMBA–TPA-induced skin tumours in MmuPV1-colonized mice showed similar proliferative and mutational signatures to those in sham-infected mice and lacked MmuPV1 viral transcripts (Extended Data Fig. 3i–k).

To examine the effect of MmuPV1 on carcinogenesis that is driven by UV radiation, MmuPV1 back-skin infection was performed in immunocompetent SKH-1 mice (Extended Data Fig. 4a). MmuPV1-infected immune mice[8] that received a single immunosuppressive dose of ultraviolet light B (UVB; 300 mJ cm^{-2}) at three months after MmuPV1 infection developed warts[9], indicating the long-term persistence of MmuPV1 colonization of the skin (Extended Data Fig. 4b, c). To avoid immunosuppressive UV exposure, MmuPV1- and sham-infected mice were treated with DMBA a week before undergoing treatment with UVB

(100 mJ cm^{-2}) three times a week for 25 weeks. MmuPV1-infected SKH-1 mice developed significantly fewer tumours over time and had a markedly lower tumour burden at the completion of the study compared to sham-infected controls (Fig. 2a–d). A small subset of SKH-1 mice that had persistent warts two months after MmuPV1 back-skin infection[8] were vaccinated with MmuPV1 live virus particles intraperitoneally three times over a two-week period. Four weeks later, five out of nine mice developed immunity against MmuPV1, as demonstrated by the rejection of their persistent warts (Extended Data Fig. 4d). The mice with acquired immunity against MmuPV1 developed markedly fewer skin tumours compared with the non-immune mice ($P = 0.0159$; Extended Data Fig. 4d, e). We detected a significant increase in the total number of $CD8^+$ T cells and the ratio of epidermal $CD8^+$ T_{RM} cells to total T cells in the skin of MmuPV1-colonized mice compared with their sham-infected controls at the completion of the UV carcinogenesis protocol (Fig. 2e–g). Furthermore, the total numbers of T cells and $CD8^+$ T cells were markedly increased in skin tumours of MmuPV1-colonized mice (Extended Data Figs. 4f–m, 5a–c). The levels of skin and tumour-infiltrating $CD3^-CD45^+$ leukocytes and $CD4^+$ T cells were unchanged between the two groups (Extended Data Figs. 4f–m, 5d–f).

To determine the role of $CD8^+$ T cells in mediating the anti-tumour immunity induced by papillomavirus skin colonization, SKH-1 mice were infected with MmuPV1 or sham-infected with MmuPV1 virus-like particles (sham(VLP)). MmuPV1- and sham(VLP)-infected mice underwent $CD8^+$ T cell depletion, mediated by anti-CD8 antibodies, together with the UV carcinogenesis protocol (Extended Data Fig. 5g, h). Notably, MmuPV1-colonized SKH-1 mice that were treated with IgG control developed markedly fewer tumours compared to the MmuPV1-colonized mice that underwent T cell depletion, and compared with both the IgG- and anti-CD8-antibody-treated control groups that were infected with sham(VLP) (Extended Data Fig. 5i, j). Consistent with our findings in other immunocompetent strains of mice, MmuPV1-colonized $Xpc^{-/-}$ mice—which are deficient in the ability to repair UV-induced DNA mutations[10]—were protected from skin cancer compared to their sham-infected controls (Extended Data Fig. 5k–n).

To determine whether β-HPVs have a similarly protective role in human skin, we used β-HPV RNA in situ hybridization (RNA ISH) to simultaneously detect the E6/7 transcripts of 25 types of β-HPV in human tissue sections (Extended Data Fig. 6). In contrast to skin lesions from an immunosuppressed patient, expression of β-HPV RNA was largely absent in the cancer cells of a SCC from an immunocompetent patient (Fig. 3a). Expression of β-HPV RNA was significantly reduced in cancer cells compared to adjacent normal skin keratinocytes among immunocompetent and immunosuppressed patients (Fig. 3b). The skin lesions of immunosuppressed patients had significantly higher β-HPV viral transcripts compared to skin lesions and samples of normal facial skin from immunocompetent patients (Extended Data Fig. 7a–e). β-HPV DNA in situ hybridization (DNA ISH) probes for 25 types of β-HPV (Extended Data Fig. 7f) detected higher viral load in an SCC from an immunosuppressed patient compared to an SCC from an immunocompetent patient (Extended Data Fig. 8a). β-HPV viral load was reduced in cancer cells compared to the adjacent normal skin of immunosuppressed patients (Extended Data Fig. 8b), and this reduction was more pronounced in the lesions of immunocompetent patients (Extended Data Fig. 8c). The higher viral activity and load in the skin cancers of immunosuppressed patients correlated with significantly fewer tumour- and skin-infiltrating $CD8^+$ T and $CD103^+CD8^+$ T_{RM} cells in their skin cancers compared with samples from immunocompetent patients (Extended Data Fig. 9a–c). Notably, β-HPV E7 peptides activated $CD8^+$ T cells isolated from the normal facial skin of immunocompetent adults (Fig. 3c, d, Extended Data Fig. 9d). By contrast, high-risk HPV16 E7 peptides did not activate skin-derived $CD8^+$ T cells (Fig. 3c, d, Extended Data Fig. 9d).

To identify the signals that lead to papillomavirus antigen presentation to T cells after abnormal proliferation of keratinocytes, we performed RNA sequencing (RNA-seq) on skin warts, MmuPV1-infected DMBA–UV-treated skin and tumours, and sham-infected DMBA–UV-treated skin and tumours of SKH-1 mice (Extended Data Fig. 10a–c). Among the 20 genes that were upregulated in both MmuPV1-induced warts and DMBA–UV-induced tumours (from both MmuPV1- and sham-infected groups) compared with skin (also from both groups), there were several immune-related genes—including the damage-associated molecular pattern (DAMP) genes *S100a8* and *S100a9* (Extended Data Fig. 10c). In human SCCs and warts, we confirmed the induction of S100 genes compared with normal skin and with seborrheic keratosis, a benign skin growth, in which *S100A8* and *S100A9* genes were downregulated compared with normal skin (Extended Data Fig. 10d–f).

The findings presented herein reveal a previously unrecognized role for commensal HPVs in cancer development. Using the colonization of skin by papillomavirus as a model, we show that MmuPV1-infected immunocompetent mice are protected against skin cancer that is induced by chemicals or UV radiation, in a $CD8^+$ T cell-dependent manner. Although specific-pathogen-free (SPF) mice may not fully reproduce the complex microbiome of human skin, our findings strongly suggest that antiviral adaptive immune responses define the role of papillomaviruses in skin carcinogenesis. Our discovery of β-HPV-specific $CD8^+$ T cells in the normal human skin is indicative of an adaptive immunity that is primed against commensal HPVs in healthy adults at baseline. These T cells that reside in the skin can target keratinocytes with active virus during their abnormal proliferation to form a wart or a skin cancer. Accordingly, T cell-based vaccines against commensal HPVs may provide an innovative approach to boost this antiviral immunity in the skin and help prevent warts and skin cancers in high-risk populations. In addition, increasing anti-HPV immunity may improve the efficacy of immune checkpoint blockade therapy against SCC[11]. Given the emerging diversity of the skin virome[12], it is critical to characterize the viral communities that reside in the skin of immunocompetent and immunosuppressed individuals and determine how these viruses contribute to human health and disease.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1719-9.

1. Grulich, A. E., van Leeuwen, M. T., Falster, M. O. & Vajdic, C. M. Incidence of cancers in people with HIV/AIDS compared with immunosuppressed transplant recipients: a meta-analysis. *Lancet* **370**, 59–67 (2007).
2. Nehal, K. S. & Bichakjian, C. K. Update on keratinocyte carcinomas. *N. Engl. J. Med.* **379**, 363–374 (2018).
3. Wang, J., Aldabagh, B., Yu, J. & Arron, S. T. Role of human papillomavirus in cutaneous squamous cell carcinoma: a meta-analysis. *J. Am. Acad. Dermatol.* **70**, 621–629 (2014).
4. Chockalingam, R., Downing, C. & Tyring, S. K. Cutaneous squamous cell carcinomas in organ transplant recipients. *J. Clin. Med.* **4**, 1229–1239 (2015).
5. Ingle, A. et al. Novel laboratory mouse papillomavirus (MusPV) infection. *Vet. Pathol.* **48**, 500–505 (2011).
6. Howley, P. M. & Pfister, H. J. Beta genus papillomaviruses and skin cancer. *Virology* **479–480**, 290–296 (2015).
7. Hufbauer, M. & Akgül, B. Molecular mechanisms of human papillomavirus induced skin carcinogenesis. *Viruses* **9**, 187 (2017).
8. Wang, J. W. et al. Immunologic control of *Mus musculus* papillomavirus type 1. *PLoS Pathog.* **11**, e1005243 (2015).
9. Uberoi, A., Yoshida, S., Frazer, I. H., Pitot, H. C. & Lambert, P. F. Role of ultraviolet radiation in papillomavirus-induced disease. *PLoS Pathog.* **12**, e1005664 (2016).
10. Sands, A. T., Abuin, A., Sanchez, A., Conti, C. J. & Bradley, A. High susceptibility to ultraviolet-induced carcinogenesis in mice lacking XPC. *Nature* **377**, 162–165 (1995).
11. Migden, M. R. et al. PD-1 blockade with cemiplimab in advanced cutaneous squamous-cell carcinoma. *N. Engl. J. Med.* **379**, 341–351 (2018).
12. Tirosh, O. et al. Expanded skin virome in DOCK8-deficient patients. *Nat. Med.* **24**, 1815–1821 (2018).

Methods

Human tissue studies

Discarded de-identified human tissue samples were obtained through Mohs surgery clinics and the pathology department at Massachusetts General Hospital. The skin lesions and normal skin samples were processed for immune cell or RNA isolation, or obtained as formalin-fixed paraffin-embedded sections for histological assays.

Animal studies

All mice were housed under pathogen-free conditions in the animal facilities at Massachusetts General Hospital and the University of Louisville in compliance with animal care and all relevant ethical regulations. Six-to-ten-week-old female C57BL/6J mice (Jackson Laboratory; 000664), female FVB mice (Charles River; 207), female SKH-1 Elite mice (Charles River; 477) and male and female *Xpc*$^{-/-}$ mice (Jackson Laboratory; 010563) were used in the immunocompetent arms of this study. Female *Cd4*$^{-/-}$*Cd8*$^{-/-}$ mice in the FVB background were used as T cell-deficient hosts (provided by D. G. DeNardo) (*Cd8*$^{-/-}$: Jackson Laboratory; 032563). Age- and gender-matched groups of mice were used in all experiments. Wherever possible, mice were randomized into test versus control groups and power analysis was used to determine the optimal number of mice in each group. In tumour studies, the onset of skin tumours and tumour counts were recorded from the time of DMBA treatment (week 0) and the maximum tumour diameter allowed was 2 cm. MmuPV1-infected mice were housed in a biocontainment unit in an animal facility at University of Louisville in accordance with animal care regulations.

Statistics and reproducibility

A two-tailed Mann–Whitney *U* test was used for tumour counts and T cell-activation assays. A two-tailed paired *t*-test was used for comparing RNA ISH and a two-tailed Wilcoxon matched-pairs signed-rank test for comparing DNA ISH signal counts between skin cancers and their adjacent normal skin. A two-tailed unpaired *t*-test was used for immunostained cell counts, RNA ISH signal counts comparing skin lesions to normal human skin and other continuous variables. A log-rank test was used as the test of significance for the time to tumour onset outcomes and a two-tailed Fisher's exact test for skin cancer anatomical distribution outcomes. Pearson's χ^2 tests were used for other categorical variables. A P value of less than 0.05 was considered significant. All bar graphs and dot plots show either mean + s.d. or mean ± s.d., as indicated. Data are representative of at least two independent sets of experiments with similar results.

Study approval

Analysis of de-identified samples of human tissue was reviewed and approved by Massachusetts General Hospital Institutional Review Board (IRB). Massachusetts General Hospital and University of Louisville Institutional Animal Care and Use Committee (IACUC) approved the animal studies.

Purification of MmuPV1

MmuPV1 viral stock was purified from MmuPV1-induced muzzle warts of B6.Cg-Foxn1^{nu}/Foxn1^{nu} mice using the caesium chloride gradient method following a protocol described previously[13]. In brief, muzzle warts of B6.Cg-Foxn1^{nu}/Foxn1^{nu} mice were homogenized by pulverization with a mortar and pestle in liquid nitrogen, and then homogenized with a tissue grinder (DWK Life Sciences; 885450-0023). The tissue was then subjected to three freeze–thaw cycles between liquid nitrogen and a 37 °C water bath, and sonicated for two minutes (amplitude of 20, 10-s pulse). Caesium chloride (Sigma-Aldrich; 289329) dissolved in phosphate-buffered saline (PBS) was added to the wart homogenate for a final density of 1.3623 g ml^{-1}, determined using a refractometer (product discontinued). The tissue was ultracentrifuged overnight at 36,000 r.p.m and opaque bands at densities ranging from 1.27 to 1.31 g ml^{-1} were extracted. Extracted bands were dialysed three times for eight hours using Slide-A-Lyzer cassette (VWR; PI66230) in 3 l PBS. The purity of the viral preparation was confirmed using SDS–PAGE .

MmuPV1 inoculation

Back skin of the wild-type, *Xpc*$^{-/-}$ and *Cd4*$^{-/-}$*Cd8*$^{-/-}$ mice was shaved with an electric razor and waxed. Next, skin was scarified using a nail file for 10–20 passages across the skin to generate microaberrations in the skin barrier, which was accompanied by skin erythema. Purified virus inoculum (20 µl) was pipetted onto scarified skin and spread homogenously. The same viral inoculum was used for all infected mice, which yielded the development of confluent warts on the back skin of T cell-deficient FVB mice. Sham-infected mice received 20 µl sterile normal saline topically after skin aberration. Vaseline gauze (McKesson; 61-20056) was cut to fit the site of the injury and applied under a standard adhesive bandage. Meloxicam (0.5 mg kg^{-1}, Boehringer Ingelheim Vetmedica) was injected subcutaneously for pain relief and again the next day. Bandages were removed 48 h after inoculation and 200 µl sterile normal saline was injected subcutaneously to any lethargic mice.

PCR detection of MmuPV1 in mouse skin

To confirm skin colonization after MmuPV1 back-skin infection and at the completion of carcinogenesis protocols, DNA was isolated from the skin biopsies using the DNeasy Blood & Tissue Kit (Qiagen; 69506). PCR amplification of the MmuPV1 *L1* gene was performed following a previously described method[13] (primers are listed in Supplementary Table 4b).

Wart development

For ten weeks following viral infection or sham infection, mice were monitored for the development of warts. As previously described[14], mice with warts that lasted for longer than two months were considered to have 'persistent' warts. We classified these mice as 'non-immune' and they were subjected to T cell transfer or MmuPV1 vaccination before entering the chemical and UV carcinogenesis studies. Mice that showed either no wart development or spontaneous wart rejection were classified as 'immune' and entered into carcinogenesis studies. MmuPV1 vaccination in wart-bearing SKH-1 mice was performed by intraperitoneal injection of MmuPV1 virus inoculum in 200 µl sterile PBS three times over two weeks.

Isolation and transfer of T cells

MmuPV1-colonized FVB mice that never developed warts or exhibited spontaneous regression of warts by ten weeks following infection (immune mice) were used as T cell donors. A single-cell suspension of CD4$^+$ and CD8$^+$ T cells from skin-draining lymph nodes was prepared using the EasySep Mouse T Cell Isolation Kit (StemCell Technologies; 19851). To assess the MmuPV1-specific nature of T cells from MmuPV1-colonized immune mice, we transferred their sorted CD4$^+$ and CD8$^+$ T cells from skin-draining lymph nodes into *Cd4*$^{-/-}$*Cd8*$^{-/-}$ recipients. Donor mice were injected intravenously with 2 µg CD45–APC (BioLegend; 103112) three minutes before collection to exclude any circulating T cells. At collection, single-cell suspensions of skin-draining lymph nodes were stained with CD3e–PE–Cy7 (Biolegend; 100320), CD4–APC–Cy7 (Biolegend; 100414), CD8α–FITC (Biolegend; 100706) and CD62L–PerCP/Cy5.5 (Biolegend; 104432) (Supplementary Table 4a). Sorted CD45$^-$CD3$^+$CD4$^+$CD62L^{low} and CD45$^-$CD3$^+$CD8$^+$CD62L^{low} donor memory T cells[15] were injected intravenously into *Cd4*$^{-/-}$*Cd8*$^{-/-}$ mice at 129,600 cells per mouse (6:1 CD4$^+$:CD8$^+$ ratio) in 200 µl sterile normal saline. As a control for MmuPV1-specific T cells, a group of wild-type FVB mice were vaccinated against an unrelated virus (mouse parvovirus type 1) in parallel with MmuVP1-infected T cell donor mice to propagate a population of T cells that would not respond to MmuPV1. This group of T cell donors was vaccinated with a cocktail of 50 µg

polyinosinic–polycytidylic acid (poly(I:C); Sigma-Aldrich; P1530) combined with mouse parvovirus virus-like particles (VLPs) in 200 µl sterile normal saline delivered by subcutaneous injection at four sites (50 µl per site per vaccination) on the back skin at 30 days and 3 days before T cell transfer. A total of 200 µl of 5% imiquimod (Sigma-Aldrich; 1338313) dissolved in dimethyl sulfoxide (DMSO) and diluted in 100% EtOH (Sigma-Aldrich; 276855) was applied topically after each vaccination. T cell recipients (T cell-deficient *Cd4*$^{-/-}$*Cd8*$^{-/-}$ and wild-type FVB mice) were infected with MmuPV1 two days after T cell transfer, including mice that received T cells from parvovirus vaccine, and topical imiquimod-treated donors. Another subgroup of MmuPV1 T cell recipients, including T cell-deficient *Cd4*$^{-/-}$*Cd8*$^{-/-}$ and wild-type mice, received an injection of DMBA–TPA-induced primary SCC cells into their right flank and were monitored for tumour growth (Extended Data Fig. 2a). Mice were monitored closely for wart development in MmuPV1 infection cohorts and SCC growth in tumour cohorts for two months, including photographs and tumour-size measurements. To examine the presence or absence of T cells in the recipient mice, peripheral blood was collected from the mice three weeks after the T cell transfer. Around 2–3 drops of blood per mouse, extracted from the submandibular vein, was collected in 10 ml RBC lysis buffer (Biolegend; 420301), stained with CD3e–PE–Cy7, CD4–APC–Cy7 and CD8α–FITC, and examined by flow cytometry. One million T cells in 200 µl sterile normal saline from skin-draining lymph nodes of MmuPV1-colonized (immune) mice versus naive T cells were injected intravenously into the tail vein of wart-bearing (non-immune) wild-type FVB mice. The recipient mice were monitored for the resolution of their skin warts and their response to skin chemical carcinogenesis.

Chemical carcinogenesis

After infection and evidence of immunity to MmuPV1, C57BL/6J and FVB mice underwent a skin chemical carcinogenesis protocol[16]. All mice were shaved and seven days later received a single dose of 100 µg DMBA (Sigma-Aldrich; D3254) in 200 µl acetone on the back skin. One week later, treatments with 6 µg TPA (Sigma-Aldrich; P1585) dissolved in 200 µl acetone were initiated (three times per week for 30 weeks in C57BL/6J and two times per week for 20 weeks in FVB cohorts). Throughout the carcinogenesis protocol, tumours were counted every week and photographs were collected every other week. The final tumour burden was determined based on the total number of palpable skin lesions that had developed on the back skin of the mice.

UV carcinogenesis

Following infection and evidence of MmuPV1 immunity, SKH-1 and *Xpc*$^{-/-}$ mice underwent a UV skin carcinogenesis protocol. Mice received a single dose of 50 µg (SKH-1) or 100 µg (*Xpc*$^{-/-}$) DMBA in 200 µl acetone on the back skin. One week later, mice received narrow-band ultraviolet B (UVB) (302–312 nm) 3 times weekly for up to 25 weeks (SKH-1) or 30 weeks (*Xpc*$^{-/-}$) via a UVP Black-Ray Lamp UVB (VWR; 36575-052), which was periodically calibrated using an International Light IL1400A Digital Light Meter (International Light Technologies). Mice received 100 mJ cm^{-2} UVB at each UV treatment time point. This is considered a sub-erythemic dose for a fair-skinned individual of average tanning ability (Fitzpatrick skin types I and II), which approximates to 25–50 min of sun exposure in Florida at midday in the summer[17,18]. Throughout the carcinogenesis protocol, tumours were counted every week and photographs were collected every other week. Any palpable discrete lesion that was discontinuous or separate from other lesions was considered a tumour. Tumour counts were performed by a single individual to maintain consistency from week to week. The final tumour burden was determined on the basis of the total number of palpable discrete skin lesions that had developed on the back skin of the mice after DMBA treatment. For the immunosuppressive UV-dosing experiment, SKH-1 mice received 300 mJ cm^{-2} UVB on their back skin once.

CD8$^+$ T cell depletion

SKH-1 mice were infected with MmuPV1 or sham-infected with MmuPV1 VLPs (L1Met30; 105 µg in 40 µl PBS per mouse[19]) applied to their abraded back skin. Four weeks later, MmuPV1-infected immune mice and sham(VLP)-infected controls were started on anti-CD8 (rat anti-mouse CD8α; BioXCell; YTS 169.4) or IgG (rat isotype control; Sigma-Aldrich) antibody treatment at 750 µg in 200 µl sterile PBS (first dose) followed by 250 µg in 200 µl sterile PBS weekly by intraperitoneal injection (Extended Data Fig. 5g, Supplementary Table 4a). One day after the first antibody treatment, mice underwent the UV carcinogenesis protocol as described above.

Hras-mutation-specific PCR

After the carcinogenesis protocol, DNA was extracted from tumours and skin of MmuPV1-infected, sham-infected or untreated wild-type FVB mice using the DNeasy Blood & Tissue Kit (Qiagen; 69506). Mutation-specific primers were designed as previously described with the addition of a wild-type specific primer[20] (primers are listed in Supplementary Table 4b). PCR was performed using 500 ng of genomic DNA, 12.5 pmol of each primer, 2.5 µl 10X Klentaq1 Reaction Buffer (DNA Polymerase Technology; RB20), 200 mM dNTPs (Bio Basic; DD0056), 2.0% (v/v) DMSO, 1.25 units of Klentaq-LA (DNA Polymerase Technology; 110) and water to a final volume of 25 µl. Amplification was performed as described previously[20]. In brief, DNA was denatured at 95 °C for 5 min, then cycled 30 times through denaturation at 95 °C for 1 min, hybridization at 55 °C for 1 min and extension at 72 °C for 1 min. After cycling, extension was continued for 5 min at 72 °C. PCR products (110 bp) were analysed on a 2% agarose gel (Genesee Scientific; 20-102QD) and visualized with ethidium bromide.

Histology and immunofluorescence staining

Samples of mouse tissue were collected and fixed in 4% paraformaldehyde (PFA; Sigma-Aldrich; P6148) overnight at 4 °C. Next, tissues were dehydrated in ethanol, processed and paraffin-embedded. Sections (5 µm) of paraffin-embedded tissues from mice and humans were cut, deparaffinized and stained with H&E. For immunofluorescence staining, rehydrated tissue sections were permeated with 1× PBS supplemented with 0.2% v/v Triton X-100 (Thermo Fisher Scientific; BP151) for 5 min. Antigen retrieval was performed in antigen unmasking solution (Vector Laboratories; H-3300) using a Cuisinart pressure cooker for 20 min at high pressure. Slides were washed three times for three minutes each in 1× PBS supplemented with 0.1% v/v Tween 20 (Sigma-Aldrich; P1379). Sections were blocked with 5% (m/v) bovine serum albumin (Thermo Fisher Scientific; BP1600) and 5% (v/v) goat serum (Sigma-Aldrich; G9023). The slides were stained overnight at 4 °C with primary antibodies (Supplementary Table 4a). The following day, slides were washed as above and incubated for two hours at room temperature with secondary antibodies conjugated to fluorochromes (Supplementary Table 4a). After washing as above, slides were incubated with 1:4,000 DAPI (Invitrogen; D3571) for five minutes at room temperature, then washed as above. Slides were mounted with Prolong Gold Antifade Reagent (Invitrogen; P36930). Once stained, ten randomly selected images of the tissue at 200× magnification (that is, HPF) were obtained for each section. Blinded manual counting of CD3$^+$, CD4$^+$, CD8$^+$, CD103$^+$ and CD45$^+$ cells was performed using the ZEN Blue 'event' tool (Zeiss). Positive cells were determined by comparing fluorescent intensity to the background, which was minimized using ZEN. Further analyses were performed based on the number of double-positive cells (for example, CD3$^+$CD8$^+$) and the number of T cell subtypes in the epidermal compartment over the total number of CD3$^+$ T cells in each image.

Serology

Using methods described previously[21], anti-MmuPV1-specific antibodies in mouse serum were detected by enzyme-linked immunosorbent assay (ELISA).

RNA and DNA in situ hybridization

RNA ISH and DNA ISH were performed on formalin-fixed paraffin-embedded human and mouse tissue sections using RNAscope probes and protocols[22] (Advanced Cell Diagnostics; Supplementary Table 1). DNA probes were generated using the sense strand of viral DNA at the same binding sites as RNA probes. We used the HybEZ hybridization system to perform RNAscope assay hybridization and incubation steps. In brief, 5-µm sections were baked in a dry oven for one hour at 60 °C and immediately deparaffinized in xylene, followed by rehydration in an ethanol series. Epitope retrieval was performed by placing the slides in RNAscope 1× Target Retrieval Reagent (Advanced Cell Diagnostics; 322000) at 102 °C for 15 min and then washed. Protease treatment was performed by adding RNAscope Protease Plus (Advanced Cell Diagnostics; 322331) to the section and incubating for 30 min at 40 °C in a HybEZ Oven II (Advanced Cell Diagnostics; 321720). After probe hybridization with target probes, preamplifier and amplifier, sections were stained with Fast RED reagent (RNAscope 2.5 HD Detection Reagents–RED; Advanced Cell Diagnostics; 322360). A counterstain of 50% haematoxylin and 0.02% ammonia water was used. Positive and negative probes were used in each assay to ensure proper controls. We used probes to an endogenous housekeeping gene peptidylprolyl isomerase B (*PPIB*) (Advanced Cell Diagnostics; 313901) and the bacterial gene *dapB* (Advanced Cell Diagnostics; 310043) as positive and negative controls, respectively. We assessed RNA ISH and DNA ISH red signals under a standard bright-field microscope at 400× magnification. Ten representative areas of skin cancer and normal skin from each slide were imaged at 400× magnification and positive RNA ISH or DNA ISH signals and keratinocyte nuclei were counted in each image in a blinded manner.

Quantitative PCR

RNA samples were extracted from human tissues that were stored in Allprotect (Qiagen; 76405) at 4 °C and flash-frozen samples stored at −80 °C. A piece of tissue (approximately 50–100 mg) was washed using sterile 1× PBS and placed into a tube containing a 5-mm TissueLyser bead, then 600 µl of RNeasy Lysis Buffer (Buffer RLT; Qiagen; 79216) and 2-mercaptoethanol was added to the sample-bead mixture. The tissue was homogenized for five minutes through mechanical manipulation. The liquid was transferred into a new tube, to which 1 ml TRIzol was added. Using standard Thermo Fisher protocols for TRIzol, the solution was mixed and centrifuged at 4 °C for ten minutes. The clear supernatant was collected and 0.2 ml chloroform was added per1 ml TRIzol solution. The mixture was centrifuged and the clear supernatant retrieved. For extraction of RNA, the Allprep DNA/RNA mini kit was used (Qiagen; 80284). The clear supernatant was added to the Allprep DNA spin column and the flow through was mixed with one volume of 70% ethanol. This solution was mixed and applied to the RNAeasy spin column, and standard methods of purification and DNase digestion were followed. RNA was quantified using a nanodrop Spectrophotometer (Nano Drop Technologies; ND-1000) and 1 µg RNA was used for the reverse-transcriptase reaction using the SuperScript III RT Kit (Thermo Fisher Scientific; 18080044). RNA (1 µg) was mixed with 0.25 mg ml^{-1} random primers, 10 mM dNTP mix and nuclease free H_2O for a total of 13 µl. This sample was then incubated at 65 °C for five minutes. A mix of diluted 1× first-strand buffer, 0.1 M dithiothreitol, 40 U µl^{-1} RNaseOUT and 200 U SuperScript III was added to the nucleotide mix. The sample was then incubated in a thermocycler. The program consisted of 5 min at 25 °C, one hour at 50 °C and 15 min at 70 °C. Following PCR, cDNA samples were diluted 1:9 using UltraPure DNase/RNase-Free Distilled Water. Of the 1:9 dilution, 3 µl was used in the total 10-µl quantitative PCR (qPCR) reaction. For forward and reverse primers (Integrated DNA Technologies; Supplementary Table 4b), 0.5 µl of 10 µM concentration was used. Five microlitres of SYBR Green master mix was used along with 1 µl UltraPure DNase/RNase-Free Distilled Water per reaction for keratin 14 and β-HPV qPCRs[23]. For other gene-expression analyses, a premixed cocktail of primers and probes was added in addition to PrimeTime Gene Expression Master Mix according to the manufacturer's instructions (Integrated DNA Technologies). The qPCR was run on LightCycler 480 II (Roche; 05015278001). qPCR products were verified by electrophoresis on a 1% agarose gel at 120 V for 60 min. Analysis of relative gene expression was performed in triplicate for each sample by comparing the test genes to *GAPDH* as the reference gene. The average relative gene expressions from the normal skin samples were used to normalize the relative gene expressions in SCCs, warts and seborrheic keratoses.

RNA-seq and analysis

Total RNA was extracted from the warts of SKH-1 mice after MmuPV1 back-skin infection and skin and tumours of SKH-1 mice after completion of the UV carcinogenesis protocol using the RNeasy Mini Kit (Qiagen; 74104) according to the manufacturer's instructions. A total of 2 µg RNA per sample was used for RNA sample preparations. RNA integrity was assessed with an Agilent Bioanalyzer 2100. Libraries were prepared by Novogene using the NEBNext Ultra RNA Library Prep Kit for Illumina (New England Biolabs; E7770). Sequencing was performed by Novogene using the Illumina NovaSeq 6000 System. Reads were aligned to the mouse reference genome (mm10) using STAR. Differential expression analysis was performed by Novogene using the DESeq2 R package. Unsupervised clustering was performed and visualized as principal component analysis (PCA) and volcano plots. Original data are available in the NCBI Gene Expression Omnibus (GEO) with accession number GSE128476.

Isolation of human T cells and peptide stimulation

T cells were isolated from human skin as previously described[24]. In brief, de-identified normal facial skin samples generated as part of Mohs surgery repair were obtained. Subcutaneous fat tissue was removed from the skin tissue, and the remaining tissue was minced. Small fragments of tissue were digested in RPMI 1640 including 1% DNase-I (Sigma-Aldrich) and 0.2% collagenase-I (Thermo Fisher Scientific) for two hours at 37 °C. Then cells were passed through a 40-µm cell strainer and incubated in RPMI 1640 supplemented with 20% (v/v) FBS, 1% (v/v) penicillin/streptomycin, 1% (m/v) glutamine, 0.00035% (m/v) 2-mercaptoethanol and 50 U ml^{-1} human recombinant IL-2 (BioLegend). T cells from human skin were seeded in a 96-well plate and treated with a pool of 5 β-HPV E7 peptides (HPV5, HPV8, HPV9, HPV20, HPV38, 5 µg ml^{-1} of each peptide; custom peptides; JPT), pool of HPV16 E7 peptides (5 µg ml^{-1} of each peptide; PepMix HPV 16 (protein E7); JPT; PM-HPV16-E7) or 50 ng ml^{-1} phorbol-12-myristate-13-acetate plus 500 ng ml^{-1} ionomycin. Peptide pools were generated as 15-mers with an overlap of 11 amino acids across the length of E7 proteins. After 24 h of peptide exposure, cells were collected, stained with antibodies to surface markers for T cell activation (Supplementary Table 4a) and examined by flow cytometry (BD LSRFortessa X-20). Flow cytometry data were analysed using FlowJo software (Ashland).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The data that support the findings of this study are available from the corresponding author on reasonable request. RNA-seq data have been deposited to the NCBI GEO (accession number GSE128476).

13. Joh, J. et al. Molecular diagnosis of a laboratory mouse papillomavirus (MusPV). *Exp. Mol. Pathol.* **93**, 416–421 (2012).
14. Jiang, R. T. et al. Spontaneous and vaccine-induced clearance of Mus musculus papillomavirus 1 infection. *J. Virol.* **91**, e00699-17 (2017).
15. Chen, B. J., Cui, X., Sempowski, G. D., Liu, C. & Chao, N. J. Transfer of allogeneic CD62L$^-$ memory T cells without graft-versus-host disease. *Blood* **103**, 1534–1541 (2004).

16. Abel, E. L., Angel, J. M., Kiguchi, K. & DiGiovanni, J. Multi-stage chemical carcinogenesis in mouse skin: fundamentals and applications. *Nat. Protocols* **4**, 1350–1362 (2009).
17. D'Orazio, J. A. et al. Topical drug rescue strategy and skin protection based on the role of *Mc1r* in UV-induced tanning. *Nature* **443**, 340–344 (2006).
18. Technology Planning and Management Corporation. *Report on Carcinogens Background Document for Broad-Spectrum Ultraviolet (UV) Radiation and UVA, and UVB, and UVC* (Technology Planning and Management Corporation, 2000).
19. Joh, J. et al. MmuPV1 infection and tumor development of T cell-deficient mice is prevented by passively transferred hyperimmune sera from normal congenic mice immunized with MmuPV1 virus-like particles (VLPs).*Exp. Mol. Pathol*. **100**, 212–219 (2016).
20. Nelson, M. A., Futscher, B. W., Kinsella, T., Wymer, J. & Bowden, G. T. Detection of mutant Ha-*ras* genes in chemically initiated mouse skin epidermis before the development of benign tumors. *Proc. Natl Acad. Sci. USA* **89**, 6398–6402 (1992).
21. Joh, J., Jenson, A. B., Ingle, A., Sundberg, J. P. & Ghim, S. J. Searching for the initiating site of the major capsid protein to generate virus-like particles for a novel laboratory mouse papillomavirus. *Exp. Mol. Pathol*. **96**, 155–161 (2014).
22. Wang, F. et al. RNAscope: a novel in situ RNA analysis platform for formalin-fixed, paraffin-embedded tissues. *J. Mol. Diagn*. **14**, 22–29 (2012).
23. Dang, C. et al. E6/E7 expression of human papillomavirus types in cutaneous squamous cell dysplasia and carcinoma in immunosuppressed organ transplant recipients. *Br. J. Dermatol*. **155**, 129–136 (2006).
24. Watanabe, R. et al. Human skin is protected by four functionally and phenotypically discrete populations of resident and recirculating memory T cells. *Sci. Transl. Med*. **7**, 279ra39 (2015).

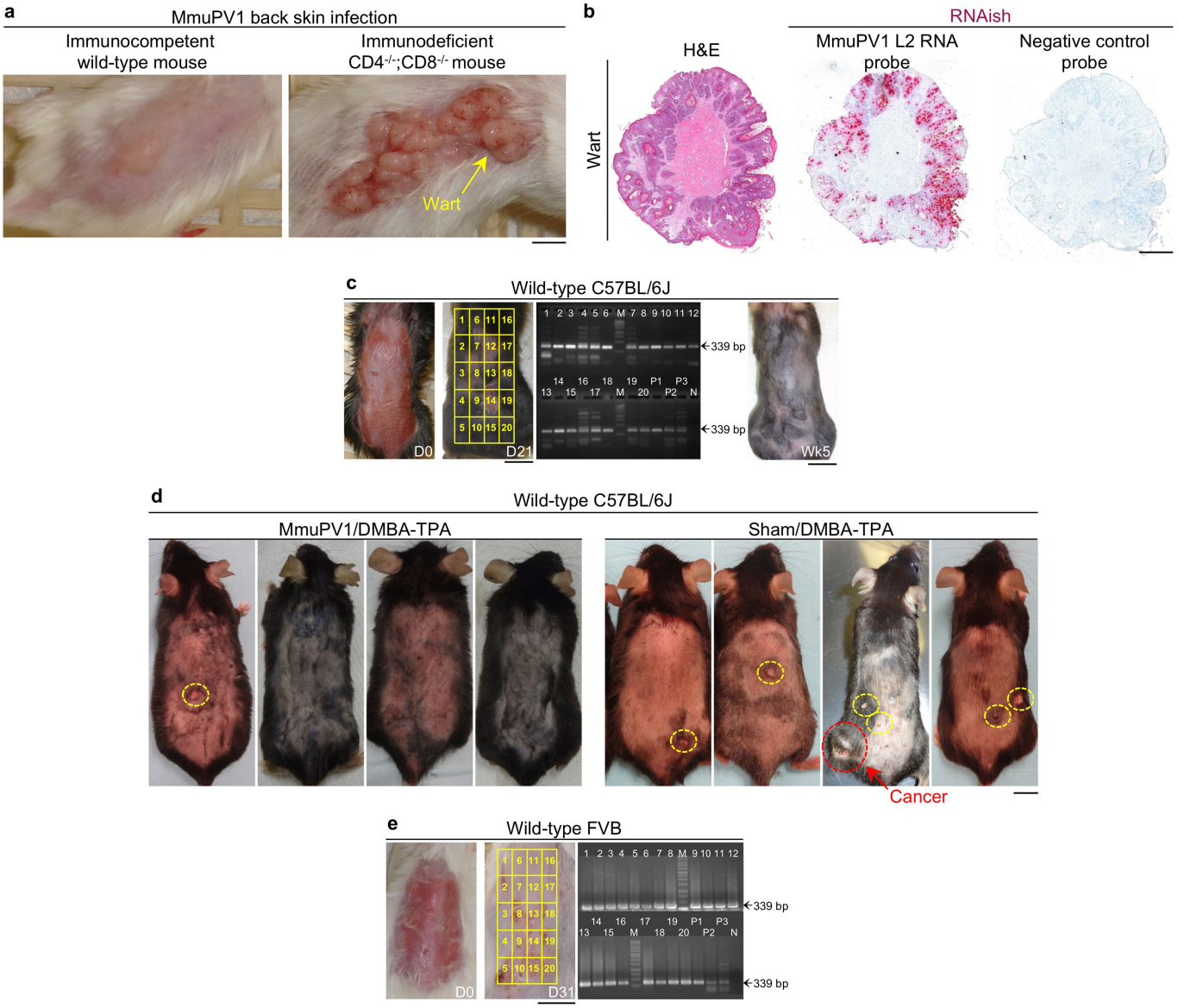
Acknowledgements We thank D. G. DeNardo for providing *Cd4*$^{-/-}$*Cd8*$^{-/-}$ mice. S.D. holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund and was supported by a Cancer Research Institute CLIP Grant; T.H. was supported by Shiseido; J.D.S., J.L.M., T.H., D.T.H., P.A.B., K.H.N. and S.D. were supported by grants from the Burroughs Wellcome Fund, Sidney Kimmel Foundation, Cancer Research Institute and NIH (K08AR068619, DP5OD021353 and U01CA233097); H.W.N., A.B.J., J.J. and the mouse studies were supported by anonymous donations for Oral Cancer Research, the Elsa U. Pardee Foundation, the Helmsley Trust Foundation and the James Graham Brown Cancer Center, University of Louisville.

Author contributions S.D. conceived the study; J.D.S., J.L.M., M.E.A., T.H., J.J. and S.D. designed the experiments; J.D.S., J.L.M., T.L., D.T.H., H.W.N., P.A.B. and K.H.N. performed the mouse experiments including tumour studies and molecular and histological analyses; M.E.A. and T.L. designed and performed RNA ISH and DNA ISH assays on human tissues; T.L. and T.H. performed immunostaining experiments on human tissue samples; T.H. performed the ex vivo CD8^{+} T cell-activation assay; J.D.S., J.L.M., M.E.A., T.L., T.H., A.B.J., J.J. and S.D. analysed and interpreted the data; J.D.S., J.L.M. and S.D. wrote the manuscript; M.M.A., R.M.N. and V.A.N. contributed clinical samples; R.M.N. determined the histological diagnosis of the human skin cancers.

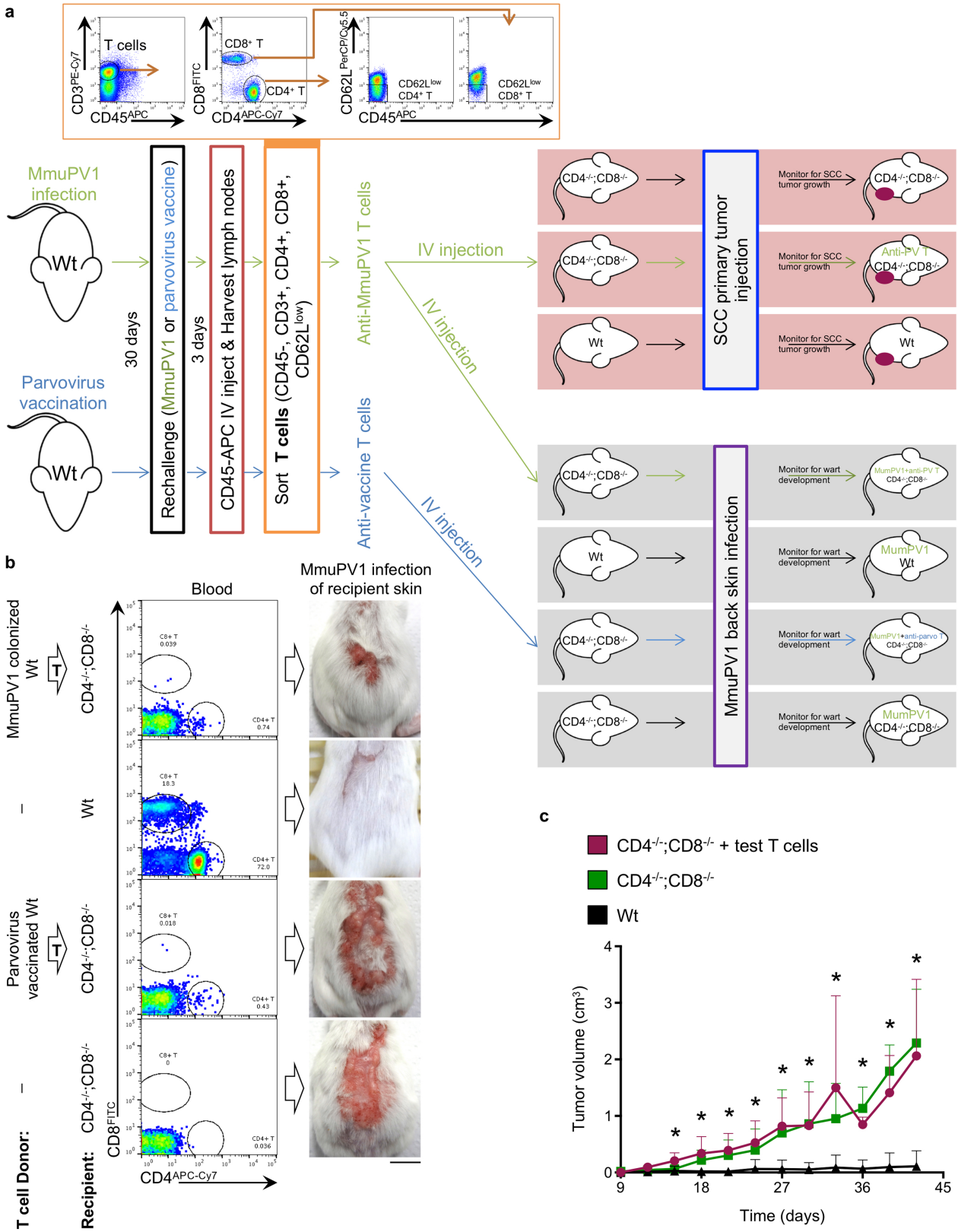
Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1719-9.
Correspondence and requests for materials should be addressed to S.D.
Peer review information *Nature* thanks Akiko Iwasaki, Thomas Tüting and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.

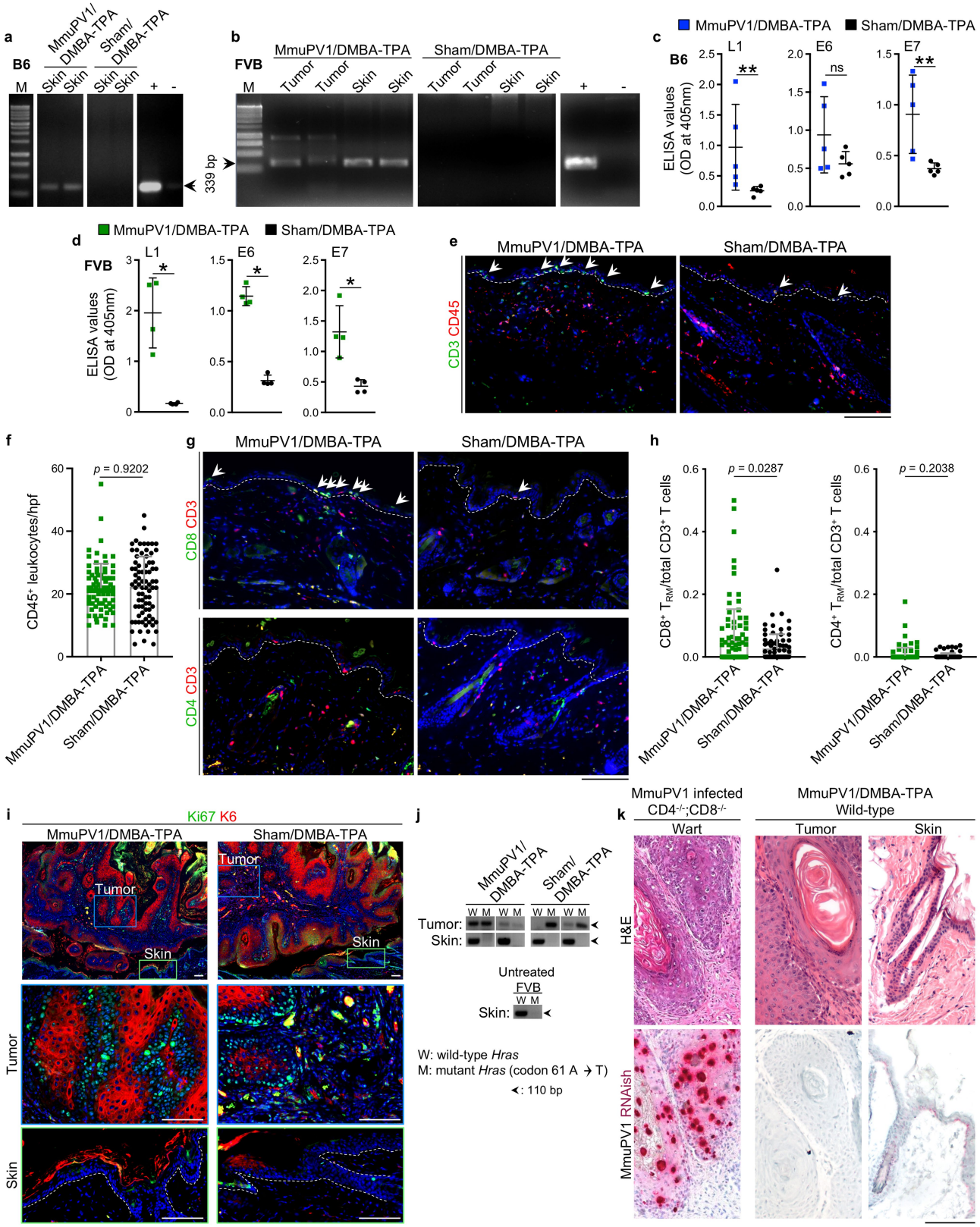


Extended Data Fig. 1 | Infection of back skin with MmuPV1 in wild-type and T cell-deficient mice and the effect of MmuPV1 colonization on the outcomes of chemical carcinogenesis in wild-type C57BL/6J mice. a, Wart burden in *Cd4*$^{-/-}$ *Cd8*$^{-/-}$ mice (right) compared with the absence of warts in wild-type mice (left) after infection of back skin with MmuPV1, at 10 weeks after infection. Note the confluent pattern of wart development in the T cell-deficient mouse. **b**, MmuPV1-induced wart in a *Cd4*$^{-/-}$ *Cd8*$^{-/-}$ mouse, stained with H&E (left), MmuPV1*L2* RNA ISH probe (middle) and negative-control RNA ISH probe (right). **c**, Left, representative images of the back skin of wild-type C57BL/6J mice on the day of MmuPV1 infection and 21 days after infection. Middle, MmuPV1*L1* PCR on 20 segments of the back skin. MmuPV1*L1* PCR bands are marked by arrows; PCR amplicon size, 339 bp. PCR primers, forward: GAGCTCTTTGTTACTGTTGTC; reverse: ATCCTCTCTTTCCTTGGGC. M, molecular-weight size marker; N, negative control; P1–P3, positive controls. Right, a typical wild-type C57BL/6J mouse five weeks after infection, highlighting the absence of warts, which was the case for 100% of the mice. **d**, Representative macroscopic images of wild-type C57BL/6J mice that were either infected with MmuPV1 on their back skin or sham-infected, and treated with DMBA–TPA. Papillomas and invasive skin cancers are highlighted with yellow and red circles, respectively. **e**, Left, Representative images of the back skin of wild-type FVB mice on the day of MmuPV1 infection and 31 days after infection, and MmuPV1*L1* PCR on 20 segments of the back skin. Mice were shaved for visualization of the skin and skin tumours. Scale bars, mouse, 1 cm (**a**, **c**–**e**); tissue, 1 mm (**b**).



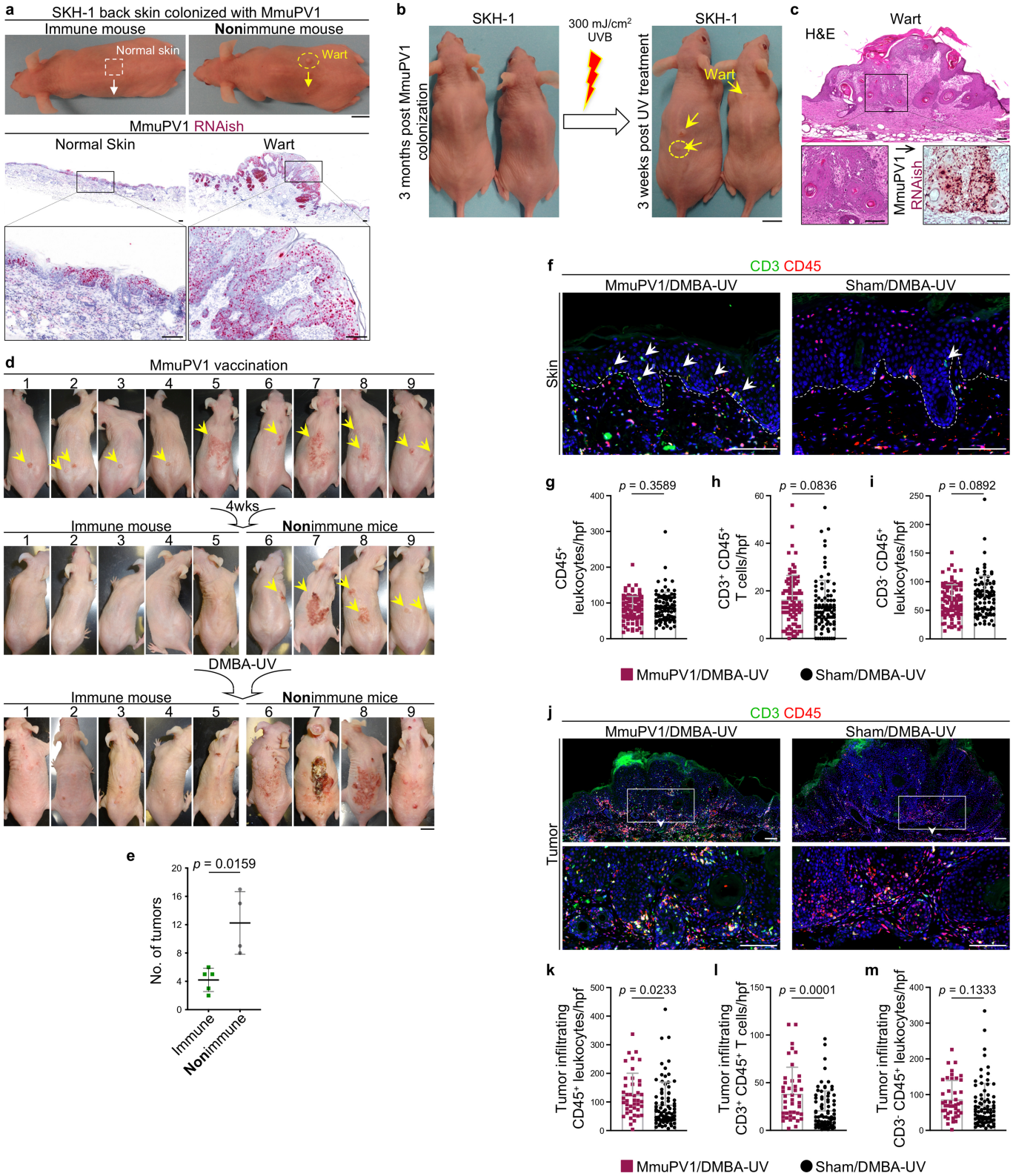
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | T cells transferred from wild-type MmuPV1-colonized immune mice to T cell-deficient mice reduce the burden of warts in mice infected with MmuPV1, but have no effect on the growth of uninfected SCC cells. a, Schematic of the T cell-transfer experiment. The inset shows the gating strategy for flow cytometry that was used to select memory T cells. T cell-donor mice received CD45–APC intravenously (IV) three minutes before the collection of lymph nodes to label and exclude the circulating immune cells. Note that the control experiment in which mice were vaccinated with parvovirus was done in parallel with the MmuPV1 challenge, and the SCC primary tumour growth experiment was done in parallel with the infection of back skin with MmuPV1. **b**, Right, representative images of the warts on the back skin of mice three weeks after infection with MmuPV1. Scale bar, 1 cm. Left, flow cytometry demonstrates the presence of $CD4^+$ and $CD8^+$ T cells in the peripheral blood of the recipient mice, indicating successful adoptive transfer of T cells ($n = 4$ per group). Wt, wild type. **c**, Growth of subcutaneously injected DMBA–TPA-induced primary SCC tumour cells in wild-type mice ($n = 9$), $Cd4^{-/-}Cd8^{-/-}$ mice ($n = 5$) and $Cd4^{-/-}Cd8^{-/-}$ mice that received T cells from MmuPV1-immune donors ($Cd4^{-/-}Cd8^{-/-}$ + test T cells) ($n = 4$). Two-tailed Mann–Whitney U test; $^*P < 0.05$ compared with the wild-type group. Data are mean + s.d.



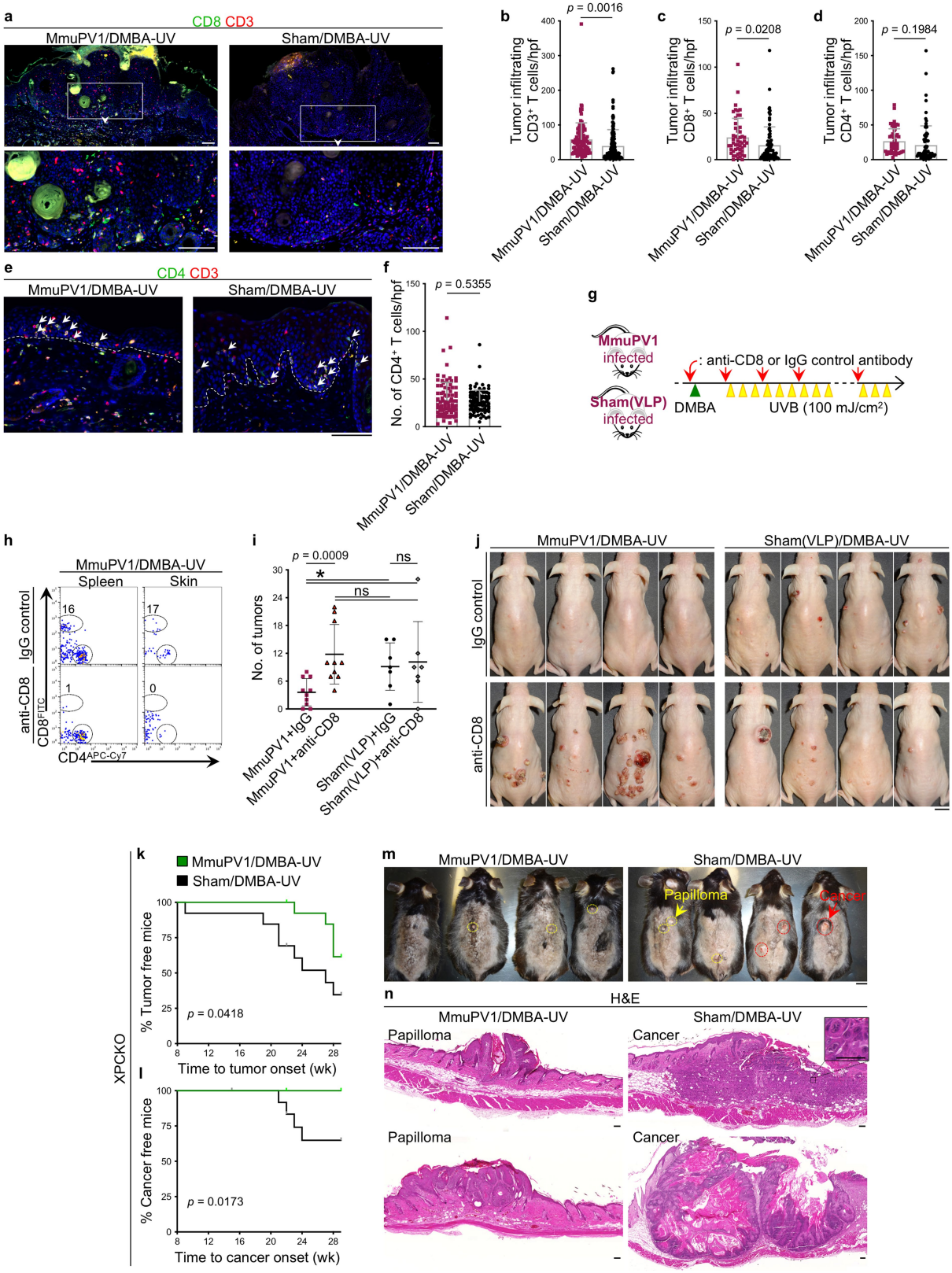
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Evidence of MmuPV1 colonization and T cells homing to the epidermis of MmuPV1-infected mice at the completion of the chemical carcinogenesis protocol. a, **b**, MmuPV1 *L1* PCR on DNA isolated from the skin of wild-type C57BL/6J (B6) (**a**) and the skin and tumour of wild-type FVB (**b**) mice more than 6 months after MmuPV1 infection. MmuPV1 *L1* PCR bands are highlighted by arrows; PCR amplicon size: 339 bp. Plus sign indicates positive control; minus sign, negative control. **c**, **d**, Anti-MmuPV1 seroconversion in DMBA–TPA-treated cohorts of C57BL/6J mice (**c**; $n = 5$ per group) and FVB mice (**d**; $n = 4$ per group). Two-tailed Mann–Whitney *U* test; $^{*}P < 0.05$, $^{**}P < 0.01$, NS, not significant. **e**, Representative images of CD3/CD45-stained skin from MmuPV1/DMBA–TPA FVB mice compared with sham/DMBA–TPA controls at the completion of the chemical carcinogenesis protocol. Arrows indicate T cells in the epidermis; dashed lines highlight the epidermal basement membrane. **f**, $CD45^{+}$ leukocytes quantified in skin sections of MmuPV1/DMBA–TPA and sham/DMBA–TPA FVB mice across ten randomly selected HPF images of normal skin per mouse and averaged across the mice in each group (two-tailed unpaired *t*-test; $n = 8$ per group). Each dot represents the leukocyte count in one high-power image. **g**, **h**, Homing of T cells to the epidermis in MmuPV1/DMBA–TPA skin compared with sham/DMBA–TPA control skin of wild-type FVB mice. **g**, Representative images of CD8/CD3- and CD4/CD3-stained skin sections. Arrows indicate epidermal $CD8^{+}$ T_{RM} cells; dashed lines highlight the epidermal basement membrane. **h**, The ratio of epidermal $CD8^{+}$ T_{RM} and $CD4^{+}$ T_{RM} cells to total $CD3^{+}$ T cells in the skin per HPF image (two-tailed unpaired *t*-test). T cells in up to ten randomly selected HPF images of normal skin per mouse were counted. Each dot represents one high-power image. $n = 10$ (MmuPV1/DMBA–TPA); $n = 9$ (sham/DMBA–TPA). **i**, Representative skin tumours from MmuPV1/DMBA–TPA and sham/DMBA–TPA wild-type FVB mice stained with keratin 6 (K6; a marker for epidermal hyperplasia) and Ki67 (a proliferation marker). Dashed lines highlight the epidermal basement membrane in the skin. **j**, PCR amplification of the wild-type (A) and mutant (T) region of the *Hras* gene in DNA of MmuPV1/DMBA–TPA and sham/DMBA–TPA tumours and skin, and untreated skin from a wild-type FVB mouse (band size, 110 bp). The A-to-T mutation in *Hras* codon 61 highlights DMBA–TPA-induced skin tumours in MmuPV1/DMBA–TPA and sham/DMBA–TPA wild-type FVB cohorts. **k**, Matched H&E and MmuPV1 *L2* RNA ISH images of a wart from an MmuPV1-infected $Cd4^{-/-}Cd8^{-/-}$ mouse, and a skin tumour and normal skin from an MmuPV1/DMBA–TPA wild-type mouse. Note the dense and confluent RNA ISH signals in the wart from the T cell-deficient mouse. After the completion of DMBA–TPA treatment, positive MmuPV1 RNA ISH signals are detectable in the normal skin of the wild-type mouse. The skin tumour from the same mouse lacks a MmuPV1 RNA ISH signal. Stained cells were counted blindly. Data are mean ± s.d. (**c**, **d**) or mean + s.d. (**f**, **h**). Scale bars, 100 μm (**e**, **g**, **i**, **k**).



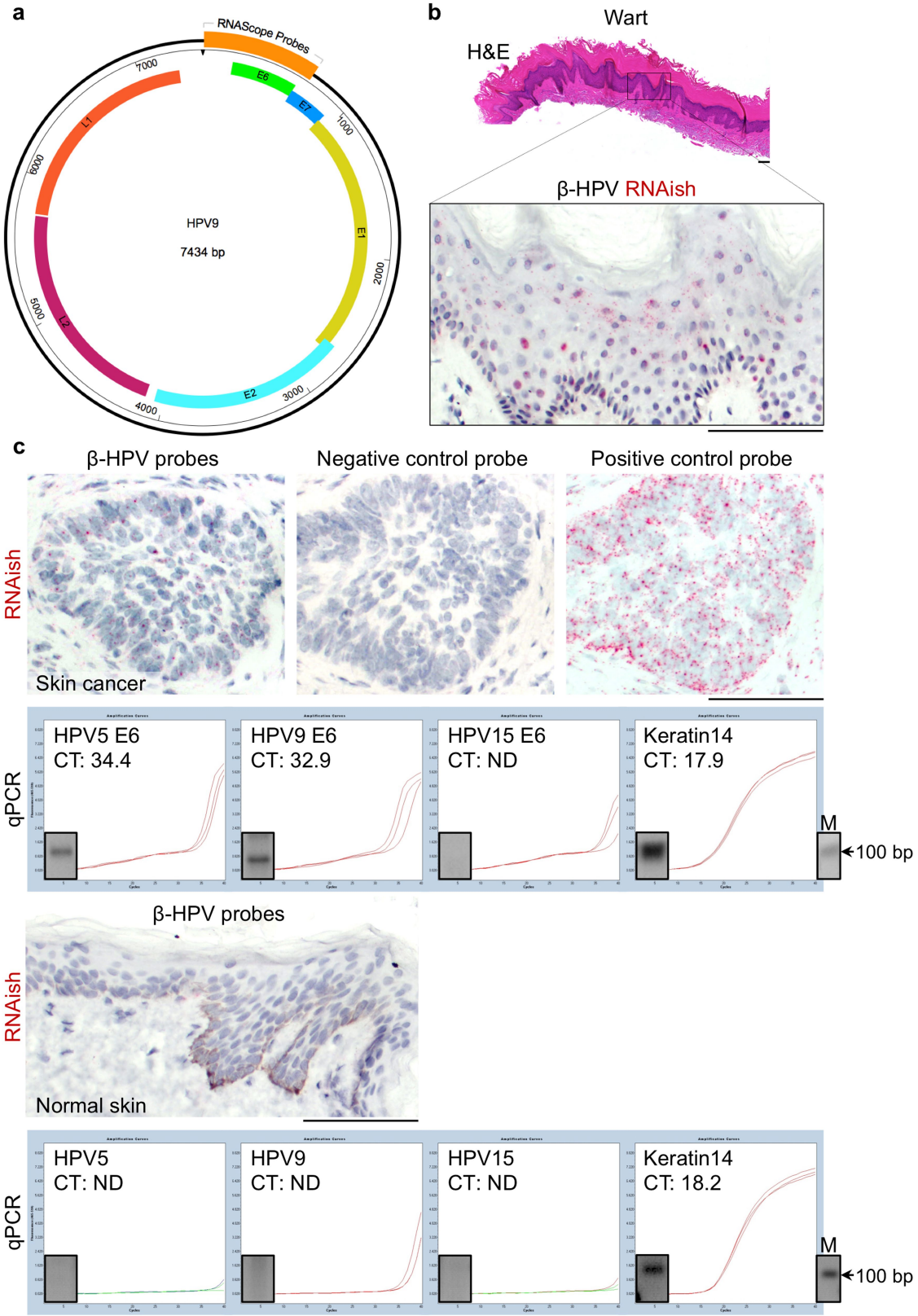
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Immunization of MmuPV1-infected SKH-1 mice with MmuPV1 vaccine protects against UV-driven carcinogenesis. a, Top, representative images of SKH-1 mice with no evidence of disease following infection (immune) and with visible warts after back-skin infection with MmuPV1 (non-immune). Bottom, MmuPV1 *L2* RNA ISH of skin from an immune and a non-immune mouse, collected three weeks after infection with MmuPV1, to detect viral activity in the normal skin and the MmuPV1-driven wart. Insets highlight the active virus in the normal skin of the immune mouse and the wart of the non-immune mouse. **b**, Macroscopic images of the SKH-1 mice three months after MmuPV1 back-skin infection. SKH-1 mice with spontaneous immunity to the virus (no wart) were treated once with an immunosuppressive dose of UVB (300 mJ cm^{-2}); images of the mice three weeks after UV treatment are shown. Arrows point to the newly developed warts on the UV-treated skin. **c**, Histological images of a wart (yellow circle), stained with H&E and MmuPV1 RNA ISH. The magnified inset highlights MmuPV1-induced cytopathic changes in the H&E image and confluent positive MmuPV1 RNA ISH signals in the wart. **d**, Macroscopic images of MmuPV1-infected SKH-1 mice that continued to have warts (yellow arrows) before MmuPV1 vaccination, four weeks after vaccination and at the completion of the UV carcinogenesis protocol. The nine wart-bearing mice were treated with MmuPV1 live virus particles intraperitoneally three times over two weeks. Four weeks later, the mice underwent the UV carcinogenesis protocol. Mice with acquired antiviral immunity ($n = 5$) are compared with non-immune mice that have persistent warts ($n = 4$). **e**, Skin tumour burden in vaccinated immune ($n = 5$) and non-immune ($n = 4$) mice treated with the UV carcinogenesis protocol. In mice with a confluent pattern of skin tumours, counts represent the individual lesions before their coalescence. Two-tailed Mann–Whitney *U* test; data are mean ± s.d. **f**, Representative images of CD3/CD45-stained skin from MmuPV1/DMBA–UV SKH-1 mice compared with sham/DMBA–UV controls at the completion of the UV carcinogenesis protocol. Arrows indicate T cells in the epidermis; dashed lines highlight the epidermal basement membrane. **g**–**i**, Skin-infiltrating total $CD45^+$ leukocytes (**g**), $CD3^+CD45^+$ T cells (**h**) and $CD3^-CD45^+$ leukocytes (**i**) quantified in CD3/CD45-stained skin sections of MmuPV1/DMBA–UV ($n = 10$) and sham/DMBA–UV ($n = 9$) SKH-1 mice across ten randomly selected HPF images of each skin sample and averaged across the mice in each group. Each dot represents one high-power image. Note the trend towards an increase in T cells and a decrease in $CD3^-$ inflammatory cells in MmuPV1/DMBA–UV skin compared with sham/DMBA–UV control. **j**, Representative images of CD3/CD45-stained cells in the skin tumours of MmuPV1/DMBA–UV SKH-1 mice compared with sham/DMBA–UV controls at the completion of the UV carcinogenesis protocol. Magnified insets highlight the immune cells in the tumour parenchyma. **k**–**m**, Tumour-infiltrating total $CD45^+$ leukocytes (**k**), $CD3^+CD45^+$ T cells (**l**) and $CD3^-CD45^+$ leukocytes (**m**) quantified in CD3/CD45-stained sections of MmuPV1/DMBA–UV and sham/DMBA–UV SKH-1 skin tumours across HPF images of each tumour and averaged across the mice in each group ($n = 12$ early skin tumours per group). Each dot represents one high-power image. Stained cells were counted blindly. Two-tailed unpaired *t*-test; data are mean + s.d. (**g**–**i**, **k**–**m**). Scale bars, mouse, 1 cm (**a**, **b**, **d**); tissue, 100 µm (**a**, **c**, **f**, **j**).



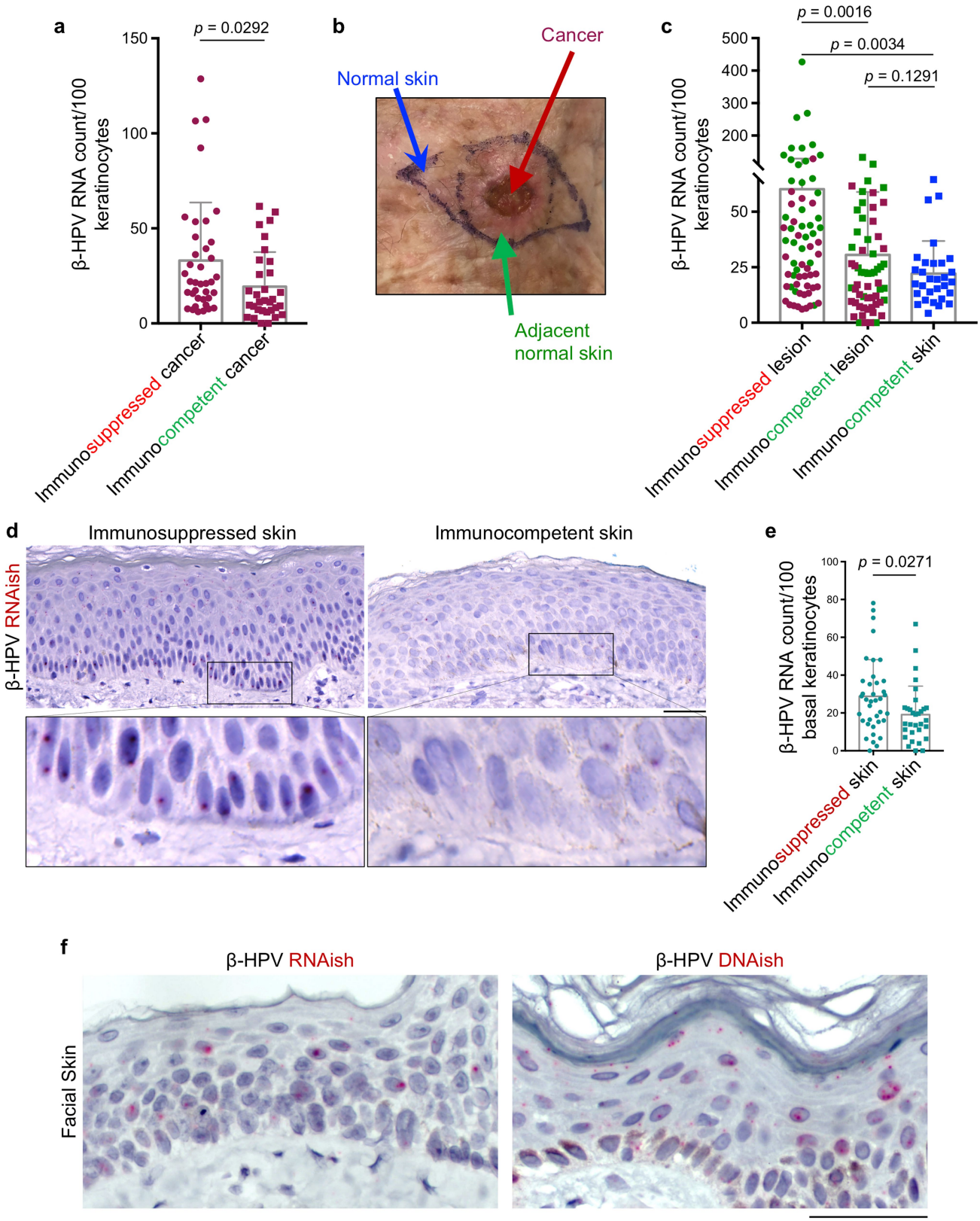
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | CD8$^+$ T cell immunity is required to protect MmuPV1-colonized mice from UV carcinogenesis and MmuPV1 colonization protects *Xpc*$^{-/-}$ mice from UV carcinogenesis. a, Representative images of CD8$^+$ T cells in the skin tumours of MmuPV1/DMBA–UV SKH-1 mice compared with sham/DMBA–UV controls at the completion of the UV carcinogenesis protocol. Magnified insets highlight T cells in the tumour parenchyma. **b**–**d**, Tumour-infiltrating CD3$^+$ (**b**), CD8$^+$ (**c**) and CD4$^+$ (**d**) T cells quantified in CD8/CD3- and CD4/CD3-stained tumour sections of MmuPV1/DMBA–UV and sham/DMBA–UV SKH-1 mice across HPF images of each tumour and averaged across the mice in each group (n = 12 early skin tumours per group). Each dot represents one high-power image. **e**, **f**, CD4$^+$ T cell infiltrates in MmuPV1/DMBA–UV and sham/DMBA–UV SKH-1 skin. **e**, Representative images of the CD4/CD3-stained skin sections. Arrows indicate epidermal CD4$^+$ T_{RM} cells; dashed lines highlight the epidermal basement membrane, **f**, Quantification of CD4$^+$ T cells per high-power image of the skin. Ten randomly selected HPF images of skin per mouse in each group are included. Each dot represents one high-power image. n = 10 (MmuPV1/DMBA–UV); n = 9 (sham/DMBA–UV). Two-tailed unpaired t-test; data are mean + s.d. (**b**–**d**, **f**). **g**, Schematic diagram of anti-CD8 or IgG antibody treatment combined with the UV carcinogenesis protocol. Four weeks after MmuPV1 or sham(VLP) infection, mice began treatment with anti-CD8 or IgG isotype control antibodies (red arrows). A day after the first treatment with antibodies, the back skin of SKH-1 mice was treated with 50 µg DMBA once (green triangle). Seven days later, mice began UVB treatment (100 mJ cm^{-2}) three times a week (yellow triangles). **h**, Flow cytometry analysis of spleen and skin of MmuPV1/DMBA–UV mice treated with anti-CD8 or IgG antibodies to evaluate the efficiency of CD8$^+$ T cell depletion at six weeks after treatment with DMBA. The percentage of CD8$^+$ T cells is shown on each plot. **i**, Skin tumour burden in MmuPV1-colonized mice treated with IgG control (MmuPV1 + IgG; n = 10) or anti-CD8 antibody (MmuPV1 + anti-CD8; n = 10), and sham(VLP)-infected mice treated with IgG control (sham(VLP) + IgG; n = 7) or anti-CD8 antibody (sham(VLP) + anti-CD8; n = 7) after DMBA–UV treatment. Two-tailed Mann–Whitney U test; *P < 0.05, NS, not significant. Data are mean ± s.d. **j**, Representative images of mice in the four treatment groups. Owing to the large skin tumours in MmuPV1-colonized CD8$^+$ T cell-depleted mice, the UV carcinogenesis study was terminated at 18 weeks after DMBA treatment. **k**, **l**, *Xpc*$^{-/-}$ (XPCKO) mice were infected with MmuPV1 on their back skin (n = 15) or sham-infected (n = 13) and subjected to the UV carcinogenesis protocol. Skin tumour outcomes are shown as the time to development of the first skin tumour (**k**) and time to development of the first invasive skin cancer (**l**) (log-rank test). Note that all *Xpc*$^{-/-}$ mice in the study were immune to MmuPV1 (that is, exhibited no wart development). **m**, Representative images of *Xpc*$^{-/-}$ mice at the completion of the 30-week UV carcinogenesis protocol. Premalignant tumours (papillomas) and invasive skin cancers are highlighted with yellow and red circles, respectively. Mice were shaved for UV treatments and the visualization of the skin tumours. **n**, Representative H&E-stained histological images of a papilloma in MmuPV1/DMBA–UV and invasive skin cancer in sham/DMBA–UV *Xpc*$^{-/-}$ mice. The inset shows the cellular atypia in the sham/DMBA–UV skin cancer (scale bar, 50 µm). Stained cells were counted blindly. Scale bars, mouse, 1 cm (**j**, **m**); tissue: 100 µm (**a**, **e**, **n**).



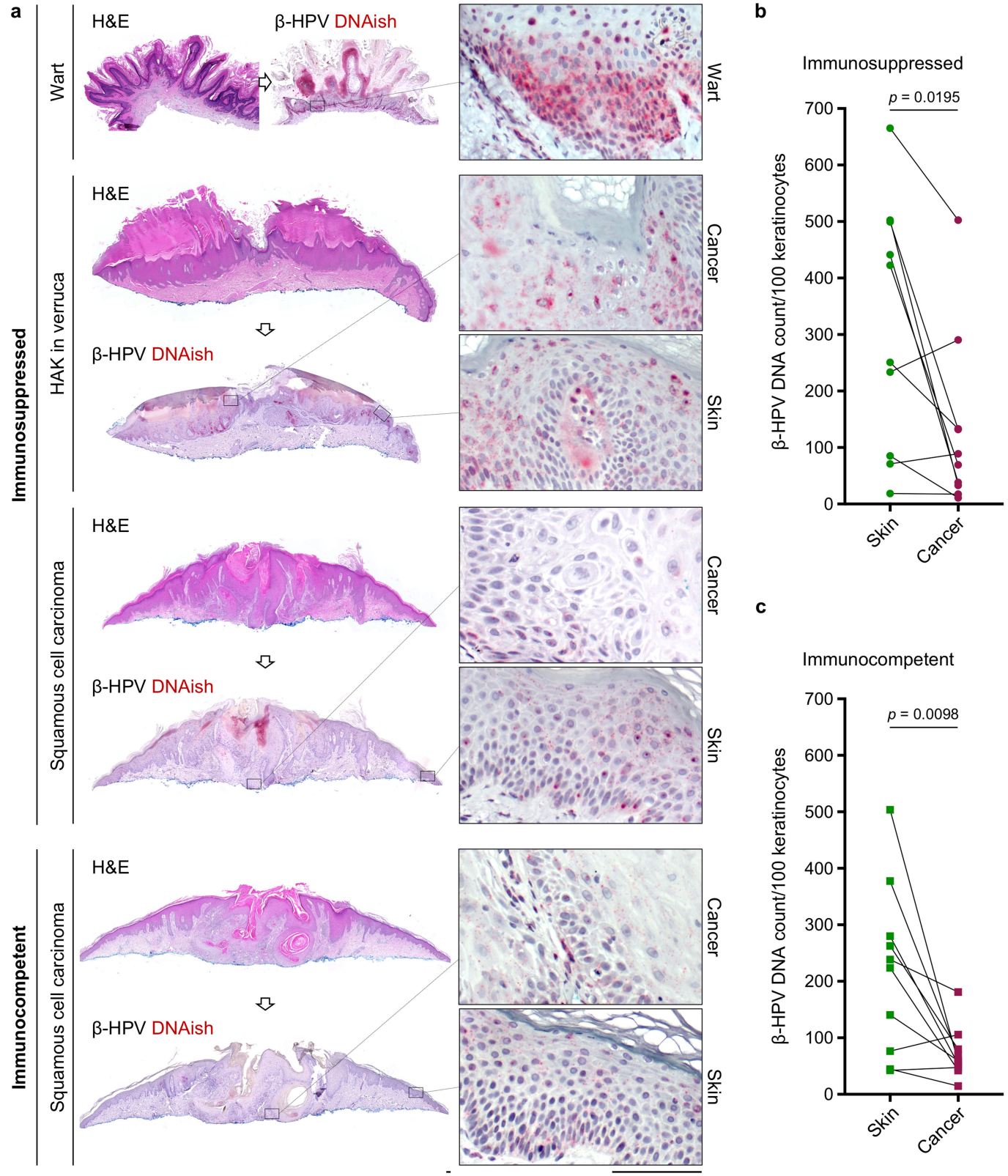
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Validation of β-HPV RNA ISH using a wart as a positive control and qPCR on RNA ISH-positive and -negative human samples. **a**, Binding site of β-HPV RNA ISH and DNA ISH probes, shown on the HPV9 genome. The RNA ISH and DNA ISH probe against each type of β-HPV comprised a pool of 20 double-Z probes that target a region of 1,000 bases (Advanced Cell Diagnostics). **b**, H&E and RNA ISH staining of a wart from a 63-year-old immunosuppressed female. Note the abundance of positive signals (red dots) throughout the wart. **c**, Top, β-HPV RNA ISH of a skin cancer from an 87-year-old immunosuppressed female, including the stains for the positive- and negative-control probes. The detection of β-HPV by RNA ISH correlates with qPCR positivity for transcripts of HPV5 and HPV9 E6 proteins in the same skin cancer. Bottom, β-HPV RNA ISH of a sample of normal skin from an 18-year-old immunocompetent African American female. The lack of RNA ISH signal (red) in this sample correlates with undetectable transcripts of HPV5, HPV9 or HPV15 E6 proteins in qPCR of the same sample. qPCR products were visualized using gel electrophoresis. PCR band sizes: HPV5 E6, 100 bp; HPV9 E6, 66 bp, HPV15 E6, 78 bp; keratin 14, 109 bp. Scale bars, 100 μm (**b**, **c**).

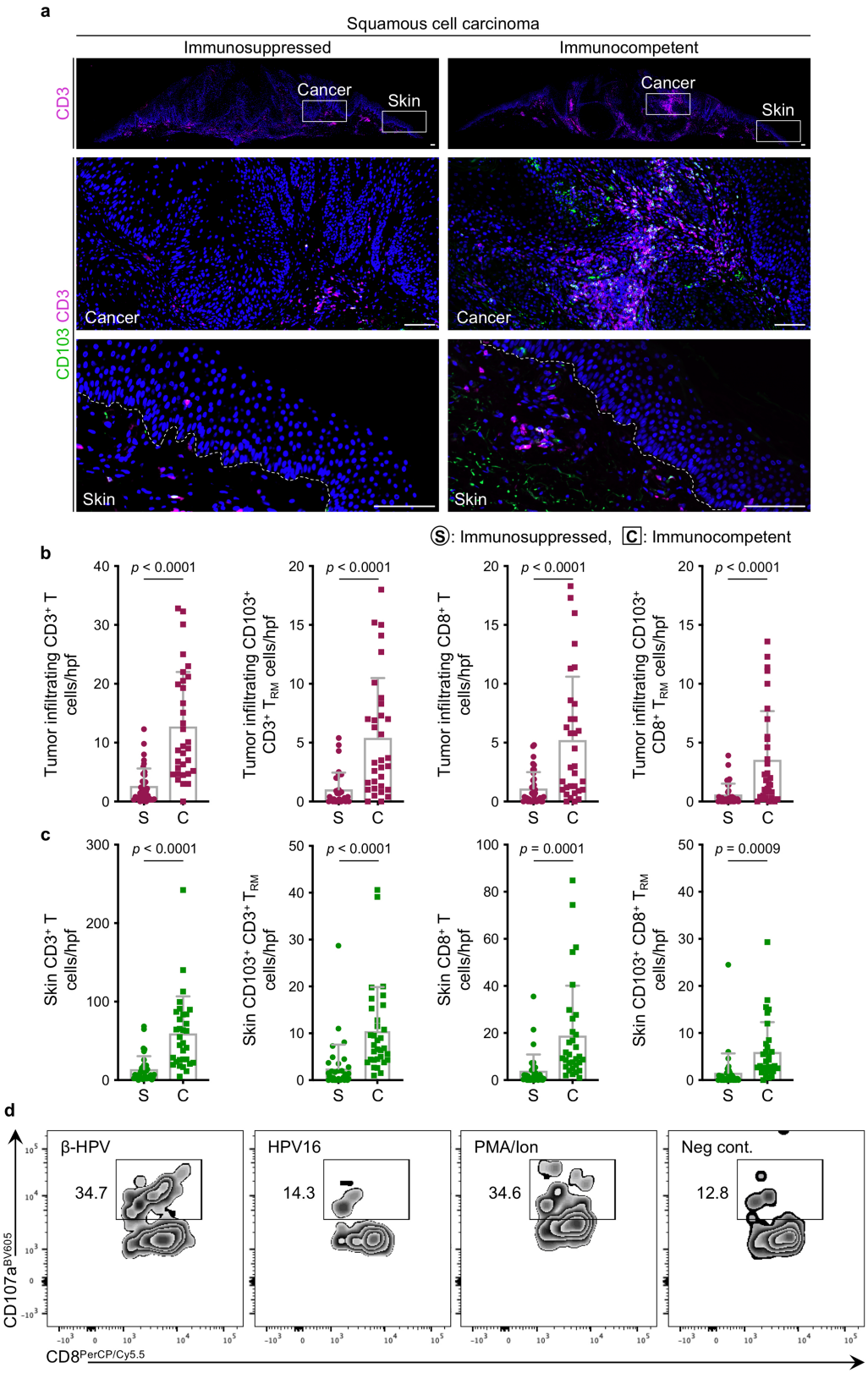


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Immunosuppressed patients have greater β-HPV viral activity in their skin lesions compared to immunocompetent patients. **a**, β-HPV RNA ISH signal counts in skin cancer cells from immunosuppressed ($n = 38$) and immunocompetent ($n = 32$) patients. **b**, Clinical image of a skin cancer surgical site showing the skin cancer (red arrow), its adjacent normal skin (green arrow) and the normal skin away from the cancer site (blue arrow). **c**, Quantification of β-HPV RNA ISH signals in high-power images across the immunosuppressed lesions, immunocompetent lesions and normal facial skin away from a cancer site. Skin lesions include β-HPV RNA ISH signal counts from skin cancer (red dots) and the adjacent normal skin (green dots). Thirty samples of normal facial skin (blue dots) from immunocompetent patients are included (18 males and 12 females; average age 71; age range 39–94). **d**, Representative low- and high-magnification images of β-HPV RNA ISH-stained normal skin samples from immunosuppressed and immunocompetent patients. Note the density and size of the apparent RNA ISH signals in basal-layer keratinocytes of an immunosuppressed patient. **e**, Density of β-HPV RNA ISH signals in basal-layer keratinocytes, quantified across 38 immunosuppressed and 31 immunocompetent skin samples. **f**, Right, β-HPV DNA ISH to detect β-HPV viral load in the skin. Compared to β-HPV RNA ISH, which marks viral transcripts, β-HPV DNA ISH detects viral load at a subcellular resolution in skin keratinocytes. Note the higher level of viral DNA ISH signals compared with RNA ISH (left), and the localization of the signals in the nucleus and cytoplasm of the keratinocytes. Two-tailed unpaired *t*-test; data are mean + s.d. (**a**, **c**, **e**). Scale bars, 50 μm (**d**, **f**).

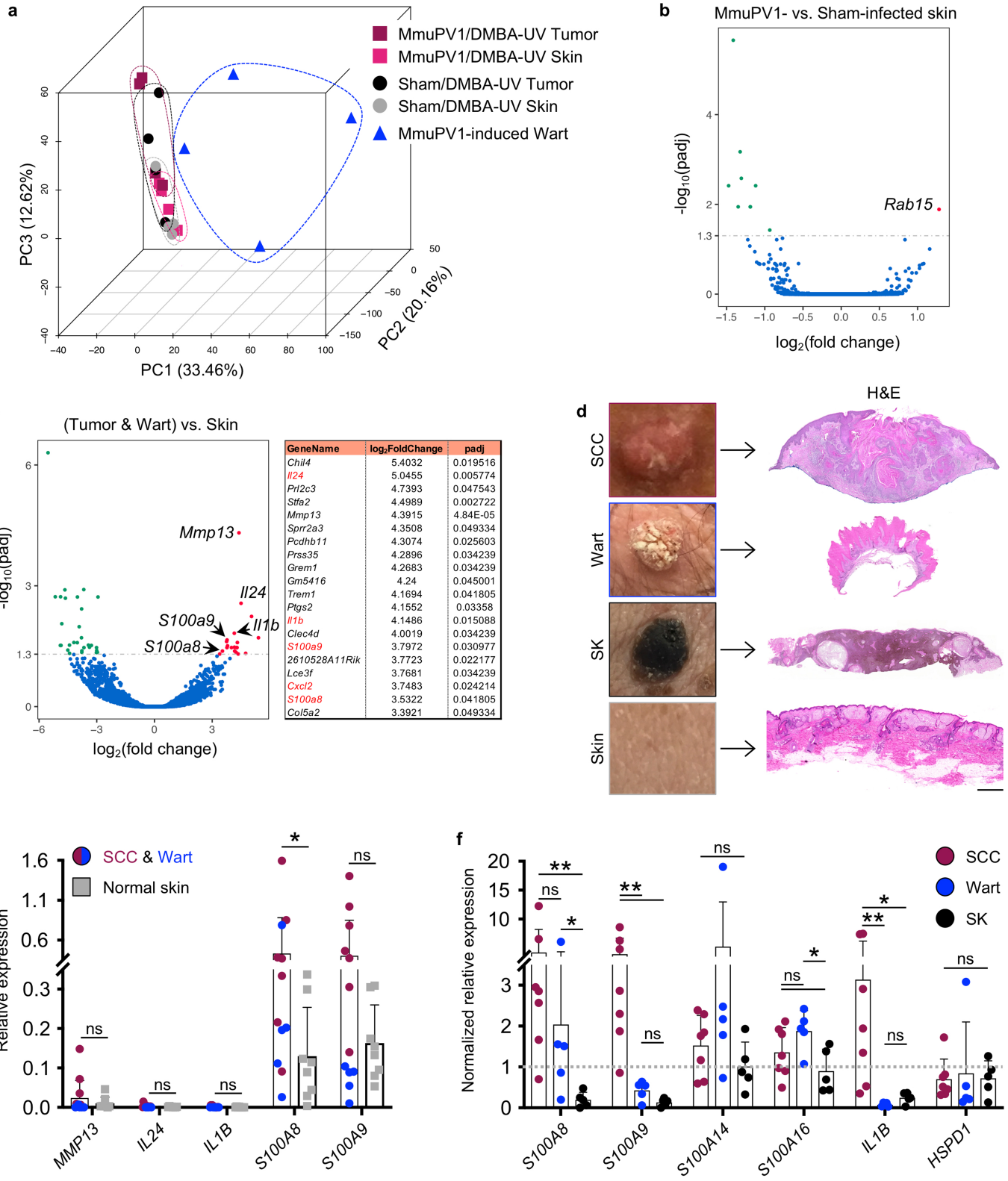


Extended Data Fig. 8 | β-HPV viral load is markedly reduced in skin cancer cells compared to their adjacent normal skin in immunocompetent patients. a, Representative DNA ISH of a wart, hypertrophic actinic keratosis arising in association with a wart (HAK in verruca), SCC in immunosuppressed patients and an SCC in an immunocompetent patient. Scale bars, 100 µm. **b**, **c**, Quantification of β-HPV DNA ISH signals in paired samples of skin cancer and the adjacent normal skin from immunosuppressed patients (**b**; $n = 10$) and immunocompetent patients (**c**; $n = 10$) (two-tailed Wilcoxon matched-pairs signed-rank test).



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Significantly fewer T and T_{RM} cells infiltrate skin cancer and the adjacent normal skin in immunosuppressed compared to immunocompetent patients. a, Representative images of CD3/CD103-stained SCC from immunosuppressed and immunocompetent patients (the same cancers are shown for β-HPV RNA ISH and DNA ISH stains in Fig. 3a and Extended Data Fig. 8a). Magnified insets highlight $CD103^+$ T_{RM} cells in the cancer and adjacent normal skin. Scale bars, 100 µm. **b**, **c**, CD3/CD8/CD103-stained sections of skin cancer were used to quantify tumour-infiltrating $CD3^+$ T, $CD103^+CD3^+$ T_{RM}, $CD8^+$ T and $CD103^+CD8^+$ T_{RM} cells infiltrating the skin cancer parenchyma (**b**), and $CD3^+$ T, $CD103^+CD3^+$ T_{RM}, $CD8^+$ T and $CD103^+CD8^+$ T_{RM} cells in the adjacent normal skin of immunosuppressed (S) versus immunocompetent (C) patients. Note that most T cells in the normal skin reside in the dermis. Stained cells were counted blindly in ten randomly selected HPF images of skin cancer and adjacent normal skin from each tissue specimen and averaged across the samples in each group; 37 immunosuppressed and 32 immunocompetent samples of skin cancer are included (skin cancer characteristics are listed in Supplementary Table 2). Each dot represents the average of the T cell counts in the high-power images from each sample. Two-tailed unpaired *t*-test; data are mean + s.d. **d**, Cytotoxic degranulation of $CD8^+$ T lymphocytes after exposure to β-HPV peptides. T cells isolated from the normal facial skin of adults were exposed to β-HPV E7 peptides (far left), HPV16 E7 peptides (middle left), PMA/ionomycin (positive control; middle right) and medium (negative control; far right). Representative flow cytometry plots are shown. The percentage of $CD107a^+CD8^+$ T cells is shown on each plot. Data represent two independent sets of experiments with similar results.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | DAMP molecules are upregulated during the development of warts and skin cancer. a, Principle component analysis (PCA) of gene-expression profiles obtained from MmuPV1-induced warts ($n = 4$; blue triangles), MmuPV1-infected skin ($n = 4$; pink squares) or sham-infected skin ($n = 4$; grey circles), and MmuPV1-infected tumours ($n = 4$; red squares) or sham-infected tumours ($n = 4$; black circles) of SKH-1 mice. Note that DMBA–UV-induced skin tumours from MmuPV1-infected mice are indistinguishable from skin tumours from sham-infected mice, whereas both have very distinct transcriptional profiles compared with MmuPV1-driven warts. **b**, **c**, Volcano plots of differentially expressed genes in MmuPV1- versus sham-infected skin (**b**; $n = 4$ per group), and skin tumours and warts ($n = 12$) versus MmuPV1- and sham-infected skin (**c**; $n = 8$). *Gm5416* is also known as *Csta3*. *P* values were calculated using the DESeq2 R package (v.21.6.3), and the resulting *P* values were adjusted using the Benjamini–Hochberg method for controlling the false discovery rate. The 20 genes that were upregulated in skin tumours and warts compared with MmuPV1- and sham-infected skin are shown in the table on the left. **d**–**f**, Analysis of the expression of immune genes in human skin lesions on the basis of the mouse RNA-seq data. **d**, Representative macroscopic and H&E-stained histological images of SCC, wart, seborrheic keratosis (SK) and unaffected human skin. Scale bar, 500 µm. **e**, Relative gene expression in SCCs ($n = 7$) and warts ($n = 5$) compared with normal skin ($n = 8$). **f**, Normalized relative gene expression in SCCs ($n = 7$), warts ($n = 5$) and seborrheic keratosis ($n = 5$) compared for several DAMP genes. Average relative gene expression in the normal skin was used for normalization. *GAPDH* is used as the reference gene. Two-tailed Mann–Whitney *U* test; $^{*}P < 0.05$, $^{**}P < 0.01$, NS, not significant; data are mean + s.d. (**e**, **f**).

nature research

Corresponding author(s): Shadmehr Demehri, M.D., Ph.D.

Last updated by author(s): Aug 16, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Cytomation software (FACS), BD FACSDiva (flow), Zeiss Zen software (histological imaging), AlphaView software (PCR gel imaging), Applied Biosystems 7500 Real-Time PCR System software (qPCR), Illumina NovaSeq 6000 (RNA Sequencing)
Data analysis	Graphpad Prism 7 and 8, Flowjo 10, Zeiss Zen Blue, DESeq2 R package (used for RNA Seq analysis)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The authors declare that the data supporting the findings of this study are available within the paper [and its Source Data Files]. RNA-seq data are deposited to the Gene Expression Omnibus (GEO). Raw data supporting the findings of this study are available from the corresponding author upon reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size were determined based on preliminary studies conducted in our laboratories and power analysis. For experiments in which no preliminary data was available in our laboratories or in similarly published research, the sample size chosen were sufficient to determine significance in all assays, with reproducible statistical significant difference between conditions in all the experiments. For animal experiments, we followed the 3 R's of animal research and used only the number of animals required to reach conclusive outcomes.
Data exclusions	No data was excluded from the study.
Replication	Animal studies were repeated in two independent sets of experiments with similar results. Representative data were repeated at least two times with similar results.
Randomization	Simple randomization scheme was applied. Mice were randomly allocated to experimental or treatment groups. Test and control mice of the same strain were gender and age matched.
Blinding	For tumor outcome measurements in animal studies, investigators were aware of host genotype and infection status of the mice while relying on unbiased measurements of quantitative parameters. The histological analyses were carried out in a blinded manner.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Immunostaining Primary Antibodies, Clone, Company:
Rat Anti-Mouse/Human, CD3, ab11089, CD3-12, Abcam, Cambridge, MA
Rabbit Anti-Mouse, CD4, ab183685, EPR19514, Abcam
Rabbit Anti-Mouse, CD8, 98941S, D4B9C, Cell Signaling Technology, Danvers, MA
Rabbit Anti-Mouse, Ki67, ab15580, Polyclonal, Abcam
Mouse Anti-Mouse, Keratin 6, ab18586, Ks6.KA12, Abcam
Rabbit Anti-Mouse, CD45, ab10558, Polyclonal, Abcam
Mouse Anti-Human, CD8a, 70306S, C8/144B Cell Signaling Technology
Rabbit Anti-Human, CD103, ab129202, EPR4166(2), Abcam

Secondary Antibodies and Staining Kits, Company:
Goat anti-Rabbit IgG, Alexa Fluor® 488 conjugate, A-11034, polyclonal, Thermo Fisher Scientific, Waltham, MA
Goat anti-Rabbit IgG, Alexa Fluor®568 conjugate, A-11036, polyclonal, Thermo Fisher Scientific
Goat anti-Mouse IgG, Alexa Fluor® 568 conjugate, P-852, polyclonal, Thermo Fisher Scientific
Goat anti-Mouse IgG, Alexa Fluor® 647 conjugate, A-21235, polyclonal, Thermo Fisher Scientific
Goat anti-Rat IgG, Alexa Fluor® 488 conjugate, A-11006, polyclonal, Thermo Fisher Scientific

Mouse Flow Cytometry Antibodies, Clone, Company:
CD3e-PE-Cy7, 100320, 145-2C11, BioLegend, San Diego, CA

CD4-APC-Cy7, 100526, RM4-5, BioLegend
CD8a-FITC, 100706, 53-6.7, BioLegend
CD45-APC, 103112, 30-F11, BioLegend
CD62L-PerCP/Cy5.5, 104432, MEL-14, BioLegend

Human Flow Cytometry Antibodies, Clone, Company:
CD3-FITC, 555916, UCHT1, BD Biosciences, San Jose, CA
CD4-APC eflour780, 47-0049-42, RPA-T4, eBioscience
CD8-PerCP/Cy5.5, 344709, SK1, BioLegend
CD45-APC, 304012, HI30, BioLegend
CD69-BV421, 310930, FN50, BioLegend
CD107a-BV605, 328633, H4A3, BioLegend
CD137-BUV395, 745737, 4B4-1, BD Biosciences

Depleting Antibodies
Anti-Mouse CD8a, BP0117, YTS 169.4 BioXCell, West Lebanon, NH
Rat Isotype Control, i4313, Polyclonal, Sigma-Aldrich, St. Louis, MO

Validation

All antibodies are commercially available and validated by previous studies done by multiple laboratories including ours. The complete descriptions and information about each antibody is provided in the corresponding Data Sheets available on the Manufacturers' website.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

Mus musculus, C57BL/6J, female, age 6-10 weeks.
Mus musculus, FVB, Wt and CD4-/-, CD8-/- female, age 6-10 weeks.
Mus musculus, SKH-1, female, age 6-10 weeks.
Mus musculus, XPC-/-, female and male, age 6-10 weeks.
Mus musculus, B6.Cg-Foxn1nu/Foxn1nu female and male, age 6-10 weeks.
All animals except CD4-/-, CD8-/- (provided by Dr. David DeNardo) were purchased from Jackson Laboratory or Charles River Laboratories.

Wild animals

The study did not involve wild animals.

Field-collected samples

The study did not involve field-collected samples

Ethics oversight

Massachusetts General Hospital and University of Louisville IACUC approved the animal studies.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics

De-identified archived skin cancers samples obtained from Massachusetts General Hospital Pathology Department. Discarded de-identified normal human skin samples were obtained through Mohs surgery clinics at Massachusetts General Hospital.

Recruitment

Normal skin and discarded tissues were collected as part of a discarded de-identified tissue sample use protocol.

Ethics oversight

Massachusetts General Hospital IRB approved the de-identified human tissue analysis.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

In experiment depicted in Fig. 1e, j and Extended Data Fig. 2, lymph nodes of MmuPV1 infected and Parvovirus vaccinated wild-

Sample preparation	type FVB mice were harvested and sorted using flow cytometry. To stain peripherally circulating lymphocytes, mice were injected intravenously with 2 ug CD45 APC three minutes prior to euthanasia. Single cell suspensions of donor lymph nodes were combined and stained for FACS. Cells were stained for 30 minutes on ice with the following surface markers diluted with PBA (500 mL PBS, 25 mL Newborn Calf Serum, 1 mL 10% sodium azide): CD3e PE-Cy7 , CD4 APC-Cy7 , CD8a FITC, CD45 APC, and CD62L PerCp/Cy5.5. Cells were sorted using MoFlo XDP, Cell Sorter. To examine the presence/absence of circulating T cells in recipient mice (Extended Data Fig. 2b), peripheral blood was collected from T cell recipient mice 2 months following the T cell transfer. 2-3 drops of blood per mouse via submandibular vein was collected in 10 mL of RBC Lysis Buffer (Biolegend, 420301), spun at 300 g for 5 min at 4C, and resuspended in a cocktail with the following antibodies in PBA: CD3e PE-Cy7 (Biolegend, 100320), CD4 APC-Cy7 (Biolegend, 100414), and CD8a FITC (Biolegend, 100706). In CD8+ T cell depletion study, a subset of MmuPV1/DMBA-UV mice that were treated with anti-CD8 or IgG isotype control were harvested 6 weeks after the start of antibody treatment. Spleen and skin single cells were prepared and stained with the following antibodies in PBA: CD45 APC, CD3e PE-Cy7 (Biolegend, 100320), CD4 APC-Cy7 (Biolegend, 100414), and CD8a FITC (Biolegend, 100706). Human skin cells were collected, used in ex vivo peptide assay and stained with antibodies to surface markers for T cell activation (Supplementary Table 5a), and examined by flow cytometry (BD LSRFortessa X-20). Flow data were analyzed using FlowJo software (Ashland, OR).
Instrument	MoFlo XDP, Cell Sorter (Beckman Coulter) using Cytomation software (for mouse studies) and BD LSRFortessa X-20 (for human flow studies)
Software	FlowJo 10
Cell population abundance	In T cell transfer experiment depicted in Fig. 1e, j and Extended Data Fig. 2, T cells from MmuPV1-immune mice (CD4: 2.53 million cells; CD8: 408,000 cells) and Parvovirus-vaccinated mice (CD4: 826,000 cells; CD8: 143,000 cells) were sorted. CD4 and CD8 cells were combined at a 6:1 ratio and 129,600 cells were transferred into each recipient mouse.
Gating strategy	For T cell transfer experiments depicted in Fig. 1e, j and Extended Data Fig. 2, memory CD4+ T cells were sorted using lymphocyte gate in FSC/SSC plot, singlets, CD3e PE-Cy7+, CD4 APC-Cy7+, CD62L PerCp/Cy5.5-, and CD45 APC- cells. Memory CD8 + T cells were sorted using CD3e PE-Cy7+, CD8 FITC+, CD62L PerCp/Cy5.5-, and CD45 APC-. Gating strategy is depicted in Extended Data Fig. 2a. T cell analyses of blood was gated on lymphocytes in FSC/SSC plot, singlets, CD3e PE-Cy7+, CD8 FITC+ vs. CD4 APC-Cy7+ T cell analyses of spleen and skin were gated on lymphocytes in FSC/SSC plot, singlets, CD45 APC, CD3e PE-Cy7+, CD8 FITC+ vs. CD4 APC-Cy7+ For the analysis of T cells isolated from human skin and used in ex vivo peptide assays, cells were gated on on lymphocytes in FSC/SSC plot, singlets, CD45 APC+, CD3e FITC+, CD4 APC eflour780+ vs. CD8 PerCP/Cy5.5+, CD69 BV421+, CD137 BUV395+, CD107a BV605+

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Article

CDK phosphorylation of TRF2 controls t-loop dynamics during the cell cycle

https://doi.org/10.1038/s41586-019-1744-8

Received: 5 March 2019

Accepted: 30 September 2019

Published online: 13 November 2019

Grzegorz Sarek[1,5], Panagiotis Kotsantis[1,5], Phil Ruis[1], David Van Ly[2,3], Pol Margalef[1], Valerie Borel[1], Xiao-Feng Zheng[4], Helen R. Flynn[1], Ambrosius P. Snijders[1], Dipanjan Chowdhury[4], Anthony J. Cesare[2] & Simon J. Boulton[1]*

The protection of telomere ends by the shelterin complex prevents DNA damage signalling and promiscuous repair at chromosome ends. Evidence suggests that the 3′ single-stranded telomere end can assemble into a lasso-like t-loop configuration[1,2], which has been proposed to safeguard chromosome ends from being recognized as DNA double-strand breaks[2]. Mechanisms must also exist to transiently disassemble t-loops to allow accurate telomere replication and to permit telomerase access to the 3′ end to solve the end-replication problem. However, the regulation and physiological importance of t-loops in the protection of telomere ends remains unknown. Here we identify a CDK phosphorylation site in the shelterin subunit at Ser365 of TRF2, whose dephosphorylation in S phase by the PP6R3 phosphatase provides a narrow window during which the RTEL1 helicase can transiently access and unwind t-loops to facilitate telomere replication. Re-phosphorylation of TRF2 at Ser365 outside of S phase is required to release RTEL1 from telomeres, which not only protects t-loops from promiscuous unwinding and inappropriate activation of ATM, but also counteracts replication conflicts at DNA secondary structures that arise within telomeres and across the genome. Hence, a phospho-switch in TRF2 coordinates the assembly and disassembly of t-loops during the cell cycle, which protects telomeres from replication stress and an unscheduled DNA damage response.

Telomere homeostasis is crucially dependent on the function of the shelterin complex but how this is regulated during the cell cycle remains uncertain. Using phospho-proteomic analysis of the shelterin complex, we identified a putative CDK2 phosphorylation site in human TRF2 at Ser365 (Ser367 in mouse; Extended Data Fig. 1a), which is abolished by treatment with λ-protein phosphatase (Fig. 1a, left and middle) or mutation of the phospho-site to alanine (Myc-tagged TRF2(S367A); Fig. 1a, right). Analysis of the cell cycle revealed that this modification is abundant in G1, G2 and M phases but is markedly reduced in S phase (Fig. 1b, Extended Data Fig. 1b).

Deletion of *Trf2* (also known as *Terf2*) results in telomere deprotection and chromosome end-to-end fusions[3] (Fig. 1c, right). By contrast, $Trf2^{F/-}$ mouse embryonic fibroblasts (MEFs) complemented with wild-type TRF2 or phospho-dead (Myc–TRF2(S367A)) or phospho-mimetic (Myc–TRF2(S367D) and Myc–TRF2(S367E)) mutants lacked telomere fusions (Fig. 1c, left, Extended Data Fig. 1c, d). The TRF2 Ser367 mutants also retained interactions with other shelterin proteins, including TRF1 and RAP1, and depletion of RAP1 did not result in telomere fusions in cells expressing Myc–TRF2(S367A)[4] (Extended Data Fig. 2a–c). Hence, TRF2 Ser367 mutants retain the ability to engage with other shelterin components and to protect telomeres against fusions.

Further analysis of TRF2-null cells expressing the TRF2 Ser367 mutants showed that the phospho-dead Ser367Ala mutant (Myc–TRF2(S367A)) resulted in high levels of telomere fragility, which indicates problems in telomere replication[5], whereas the phospho-mimetic mutants (Myc–TRF2(S367D/E)) resulted in frequent loss of telomeres, signal-free ends and high levels of extra-chromosomal telomere circles[6] (Fig. 1d–g). Because the distinct phenotypes of the phospho-dead and phospho-mimetic TRF2 Ser367 mutants resemble cells that fail to recruit the helicase RTEL1 to replication forks and telomeres, respectively[7], we reasoned that Ser365 or Ser367 of TRF2 might serve as a phospho-dependent TRF2–RTEL1 protein-interaction surface, which could cooperate with the TRFH domain that was previously shown to interact with RTEL1[8]. Pull-down experiments using biotinylated human TRF2 peptides encompassing amino acids 354–383 revealed a prominent RTEL1 band with the unphosphorylated peptide (S365) but not with the phosphorylated peptide (pS365) or an unrelated TRF2 control peptide (384–413) (Fig. 2a, b, Extended Data Fig. 3a). These results raised the possibility that the phosphorylation of Ser365 or Ser367 of TRF2 negatively regulates the interaction between TRF2 and RTEL1. The addition of λ-protein phosphatase was found to enhance this association in cell extracts (Fig. 2c), whereas addition of the phosphatase inhibitor PhosSTOP prevented a robust TRF2–RTEL1 interaction (Fig. 2c). Treatment of cells with the CDK inhibitor R-roscovitine, but not with a PLK1 inhibitor (BI-2536), also enhanced levels of Myc–TRF2 co-immunoprecipitated with RTEL1

[1]The Francis Crick Institute, London, UK. [2]Genome Integrity Unit, Children's Medical Research Institute, University of Sydney, Westmead, New South Wales, Australia. [3]School of Medicine, The University of Notre Dame Australia, Sydney, New South Wales, Australia. [4]Dana-Farber Cancer Institute, Harvard Institute of Medicine, Boston, MA, USA. [5]These authors contributed equally: Grzegorz Sarek, Panagiotis Kotsantis. *e-mail: simon.boulton@crick.ac.uk

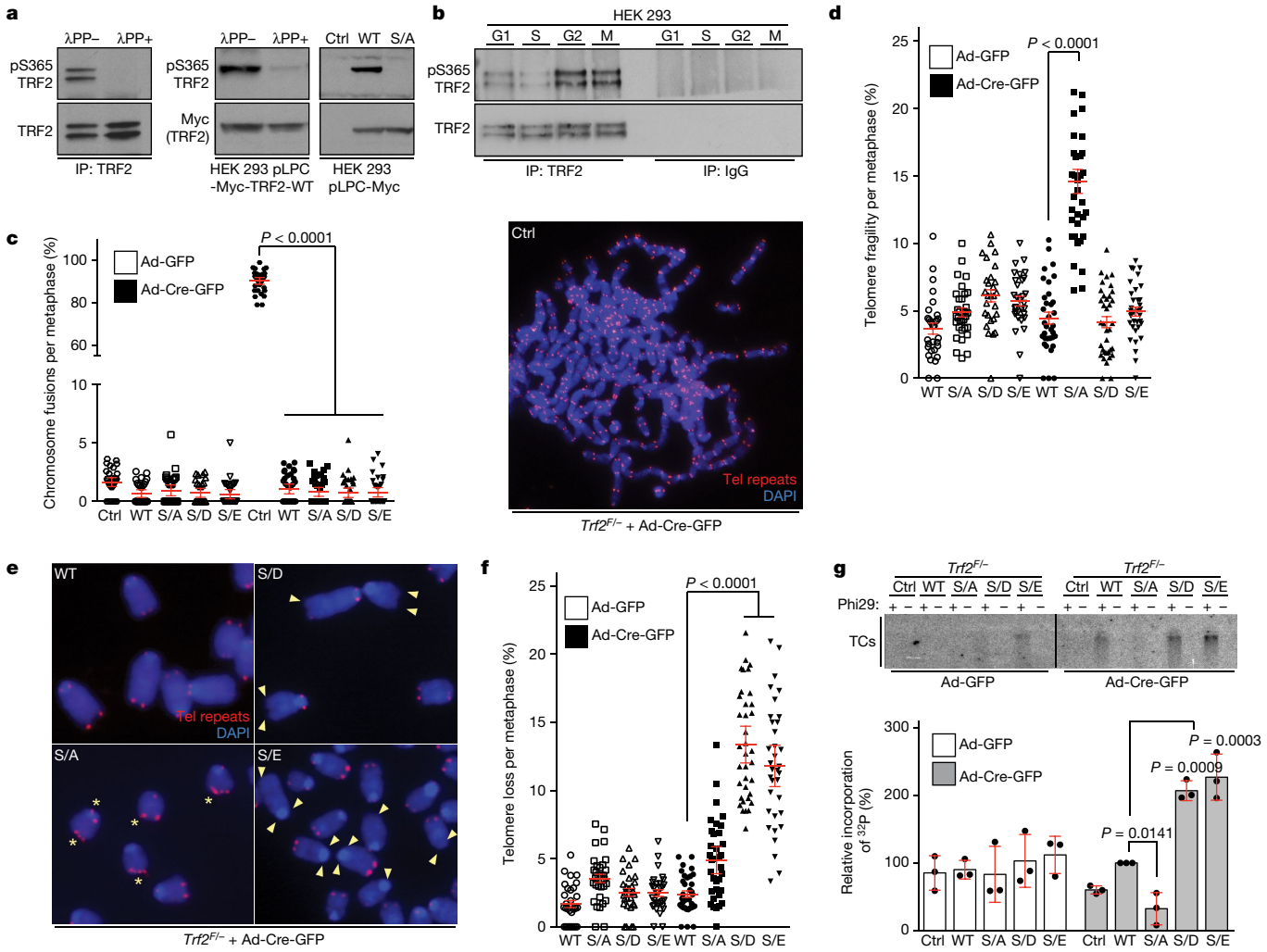


Fig. 1 | Mutations in TRF2 at Ser365 or Ser367 result in dysfunctional telomeres. a, Whole-cell extracts from HEK 293 cells (left) or HEK 293 cells expressing mouse Myc-tagged TRF2 (Myc–TRF2) or the indicated mutants (right) were pre-treated with λ-protein phosphatase (λPP) and subject to western blotting. Ctrl, control; IP, immunoprecipitate; S/A, phospho-dead mutant Myc–TRF2(Ser367Ala); WT, wild type. **b**, IgG control and endogenous phospho-TRF2 immunoprecipitates from HEK 293 cells at the indicated stages of the cell cycle. **c**–**f**, Quantification of telomere fusions (**c**), telomere fragility (**d**, **e**) and telomere loss (**f**) per metaphase from *Trf2*$^{F/-}$ MEFs stably expressing the indicated *Trf2* genotypes ($n = 35$ analysed metaphases). S/D and S/E, phospho-mimetic Myc–TRF2 mutants Ser367Asp and Ser367Glu, respectively. Data are mean ± s.e.m. Representative images of telomere FISH experiments are shown in **c** and **e**. Asterisks indicate telomere fragility, arrowheads denote loss of telomere signal. Red, telomere peptide nucleic acid (PNA) FISH; blue, DAPI. **g**, Phi29-dependent telomere circles (TCs; top) and quantification (bottom) in DNA isolated from *Trf2*$^{F/-}$ MEFs stably expressing empty vector control, wild-type or mutant TRF2, 96 h after infection with control or Cre-expressing adenovirus (Ad-GFP or Ad-Cre-GFP, respectively). Data are mean ± s.d. from three independent experiments. In **a**–**f**, experiments were independently repeated at least twice with similar results. All *P* values determined by one-way analysis of variance (ANOVA).

(Fig. 2d). The interaction of RTEL1 with wild-type Myc–TRF2, but not with the Myc–TRF2(S365A) mutant, was inhibited after incubation with recombinant cyclinA–CDK2 (Fig. 2e), which supports previous findings that Ser365 of TRF2 is a cyclinA–CDK substrate[9]. Inhibition of ERK1/2 also had no effect on the phosphorylation of TRF2(Ser365)[10] (Extended Data Fig. 3b). Whereas both phospho-mimetic mutants abolished the TRF2–RTEL1 interaction in cells, the phospho-dead Myc–TRF2(S367A) mutant interacted to a much greater extent with RTEL1 compared with wild-type Myc–TRF2 (Fig. 2f, g). Hence, TRF2 phospho-mimetic mutants abrogate the TRF2–RTEL1 interaction, which results in telomere loss and increased telomere circles, whereas the phospho-dead TRF2(S367A) mutant enhances the TRF2–RTEL1 interaction and results in telomere fragility. We conclude that CDK phosphorylation of TRF2 at Ser365 or Ser367 inhibits its interaction with RTEL1.

Given that phosphorylation of TRF2 at Ser365 is markedly reduced in S phase when TRF2 recruits RTEL1 to telomeres[8], we considered the possibility that TRF2 Ser365 is actively de-phosphorylated. TRF2 and RTEL1 complexes purified from S phase and asynchronous cells contained a number of phosphatases and/or regulatory subunits that showed increased association with TRF2 and/or RTEL1 in S phase, including UBLCP1, PP1R10, PP4R1, PP4R2, PP6R2, PP2R5C and PP6R3 (Extended Data Fig. 4a). Of these phosphatases, knockdown by short interfering RNA (siRNA) of PP4R2 or PP6R3 or their respective catalytic subunits (PP4C or PP6C, respectively), greatly reduced the TRF2–RTEL1 interaction in two different cell lines (Extended Data Figs. 4b, c, 5a, b). Co-immunoprecipitation studies confirmed that PP4R2 and PP6R3 regulatory subunits interact with TRF2 and RTEL1 in vivo (Extended Data Fig. 4d). Notably, phosphorylation of human TRF2 at Ser365 and mouse TRF2 at Ser367 were greatly enhanced after silencing of PP6R3 but not in cells subjected to PP4R2 knockdown (Extended Data Fig. 5c, d). Cells depleted of the PP4R2 or PP6R3 regulatory subunits also exhibited telomere loss (Extended Data Fig. 5e) and a greater than threefold

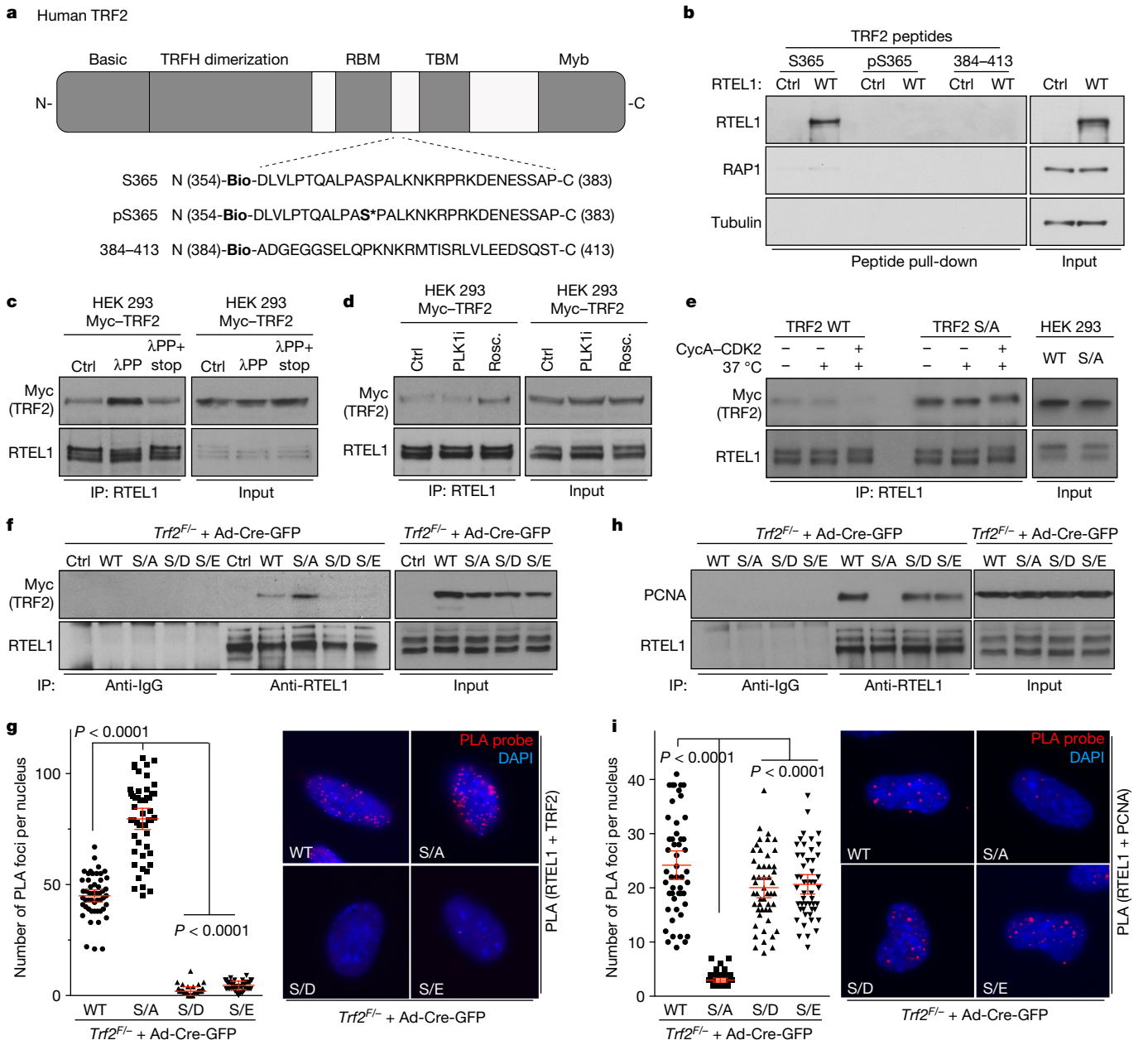


Fig. 2 | The Ser365 and Ser367 phospho-site in TRF2 controls TRF2–RTEL1 and RTEL1–PCNA interactions. a, Domain organization of mammalian TRF2 protein. TBM, TIN2-binding motif; RBM, RAP1-binding motif. **b**, Western blots of peptide pull-downs from HEK 293 cells expressing pHAGE-HA-Flag-RTEL1 (WT) or empty vector (Ctrl). **c**, Western blot of input and RTEL1 immunoprecipitates from Myc–TRF2 samples treated with vehicle control, λ-protein phosphatase or phosphatase inhibitor (λPP + stop). **d**, Western blot of input and RTEL1 immunoprecipitates from extracts of HEK 293 cells expressing Myc–TRF2 and pre-treated with vehicle, PLK1 inhibitor (PLK1i) or CDK inhibitor R-roscovitine (Rosc.) for 24 h. **e**, Immunoprecipitates from HEK 293 cells were subjected to an in vitro immunoprecipitation kinase assay with ATP and purified cyclinA–CDK2 complex (CycA–CDK2), resolved by SDS–PAGE and blotted using anti-Myc (TRF2) and anti-RTEL1 antibodies. Input (5%) is shown on the right. **f**, **h**, Immunoprecipitates were resolved by SDS–PAGE and analysed by western blotting for co-precipitated Myc–TRF2 (**f**) or PCNA (**h**). Input (5%) is shown. **g**, **i**, Cells as in **f** and **h** were quantified by in situ PLA assay for the interaction between Myc–TRF2 and RTEL1 (**g**) or PCNA and RTEL1 (**i**). Data are mean ± s.e.m.; $n = 50$ nuclei analysed. In **b**–**i**, experiments were independently repeated at least twice with similar results. *P* values determined by one-way ANOVA.

induction in telomere circles when compared with controls (Extended Data Fig. 5f). These data indicate that PP6R3 dephosphorylates TRF2 Ser365 or Ser367 to permit the transient recruitment of RTEL1 to telomeres in S phase.

Because RTEL1 facilitates global and telomere replication through its ability to interact with PCNA[11], we considered the possibility that the phospho-dead TRF2(S367A) mutant might sequester RTEL1 and limit its ability to bind PCNA. Indeed, cells expressing the phospho-dead TRF2(S367A) mutant, but not wild-type TRF2 or the phospho-mimetic mutants, were compromised for the RTEL1–PCNA interaction in co-immunoprecipitation and proximity ligation assay (PLA) experiments (Fig. 2h, i). Furthermore, analysis of global replication dynamics revealed that TRF2-null cells expressing the phospho-dead TRF2(S367A) mutant, but not wild-type TRF2 or the phospho-mimetic mutants, exhibited reduced replication fork extension rates and increased asymmetric forks across the genome (Fig. 3a, Extended Data Fig. 6a, b). Cells expressing the Myc–TRF2(S367A) phospho-dead mutant also exhibited increased levels of replication stress, which manifested

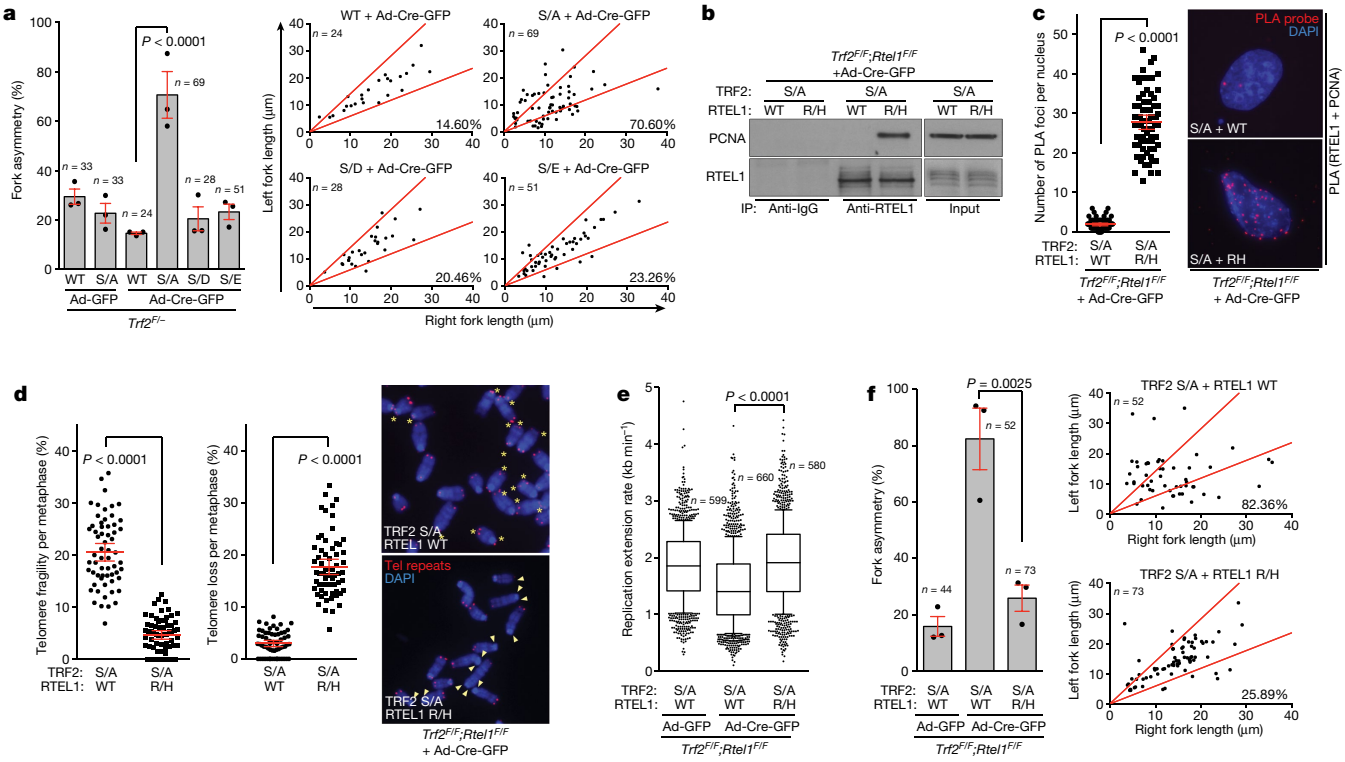


Fig. 3 | Disruption of the interaction between RTEL1 and the phospho-dead TRF2 mutant rescues abnormal genome-wide replication phenotypes. **a**, Quantification (left) and representative scatter plots (right) of fork asymmetry (*n* denotes number of analysed forks). Data are mean ± s.e.m. of triplicate experiments. **b**, Western blots of input and RTEL1 immunoprecipitates from cells of the indicated genotype. R/H, V5–RTEL1(R1237H); S/A, Flag–TRF2(S367A). **c**, Left, quantification of the PCNA–RTEL1 interaction as determined by in situ PLA assay (*n* = 70 nuclei analysed). Data are mean ± s.e.m. Right, representative images of the telomere FISH experiments. **d**, Left and middle, quantification of telomere fragility (left) and telomere loss (middle) per metaphase determined from the same cells as in **b** (*n* = 62 metaphases). Data are mean ± s.e.m. Right, representative images of the telomere FISH experiments. Asterisks denote telomere fragility; arrowheads indicate telomere loss. **e**, Quantification of global replication fork dynamics (*n* denotes the number of analysed forks). Data are mean ± s.e.m. of triplicate experiments. In box plots, horizontal line denotes the median; whiskers denote the 10th and 90th percentiles. **f**, Quantification (left) and representative scatter plots (right) of fork asymmetry (*n* denotes the number of analysed forks). Data are mean ± s.e.m. from three experiments. In **b**–**d**, the experiment was independently repeated at least twice with similar results. *P* values determined by one-way ANOVA (**a**, **e**, **f**) or unpaired two-tailed Student's *t*-test (**c**, **d**).

as micronuclei, mitotic catastrophe and increased 53BP1 nuclear foci (Extended Data Fig. 6c–f).

We reasoned that if the phospho-dead TRF2(S367A) mutant sequesters RTEL1 at telomeres, then expressing RTEL1 with a mutation in the C4C4 motif that is defective for TRF2 binding[8] (RTEL1(R1237H)) should mitigate this effect. Indeed, co-expression of the phospho-dead Flag-tagged TRF2(S367A) with the V5-tagged RTEL1(R1237H) mutant, but not with wild-type V5–RTEL1, restored the interaction between RTEL1 and PCNA in mouse ear fibroblasts (Fig. 3b, c). This co-expression also suppressed the levels of fragile telomeres (Fig. 3d), rescued the DNA replication defects (Fig. 3e, f, Extended Data Fig. 7a, b), and suppressed the formation of micronuclei, mitotic catastrophe and 53BP1 foci in mouse ear fibroblasts expressing the phospho-dead Flag–TRF2(S367A) mutant (Extended Data Fig. 7c–f). These data suggest that the TRF2(S367A) mutant sequesters the endogenous pool of RTEL1, potentially at both telomeres and pericentromeric regions[12], which restricts its ability to engage with PCNA leading to replication stress at telomeres and across the genome.

RTEL1 has been shown to unwind D-loops based on genetic studies and its ability to resolve such structures in vitro[13]. However, evidence demonstrating a direct role in unwinding t-loops in vivo, which contain a D-loop at the point of strand invasion, remains lacking. Because the TRF2(S367A) phospho-dead mutant sequesters RTEL1 at telomeres, we asked what would happen to t-loops in this context. Visualization of telomere secondary structures by Airyscan super-resolution microscopy revealed no measurable reduction in t-loop abundance in TRF2-null cells expressing wild-type TRF2[13] (Fig. 4a, Extended Data Fig. 8a). However, the frequency of t-loops was significantly diminished in TRF2-null cells expressing the Myc–TRF(S367A) phospho-dead mutant (Fig. 4b, c). Hence, sequestration of RTEL1 at telomeres leads to promiscuous t-loop unwinding, decreasing the overall levels of t-loops.

The spurious t-loop unwinding observed in the TRF2(S367A) mutant presented an opportunity to directly test whether t-loops are important for suppressing the DNA damage response (DDR) at telomeres. Analysis of TRF2-null cells expressing the phospho-dead Myc–TRF2(S367A) mutant revealed a largely ATM-dependent DDR induction at telomeres, albeit with a modest accumulation of DNA damage-induced RPA foci and activation of ATR due to telomere fragility (Fig. 4d–g, Extended Data Fig. 8b–d). Measuring the lengths of telomere contours in super-resolution micrographs revealed that linear telomeres from cells expressing Myc–TRF2(S367A) overlapped in length distribution with looped telomeres from the wild-type Myc–TRF2 control with protected chromosome ends (Extended Data Fig. 8e, f). These data suggest that promiscuous t-loop unwinding results in linear telomeres that activate an ATM-dependent DDR (Fig. 1c). Collectively, these data reveal that the t-loop structure is important for suppressing the activation of ATM at telomere ends.

In conclusion, our study identifies a phospho-switch in TRF2 that regulates the transient recruitment and release of RTEL1 from telomeres, which is required to temporarily disassemble t-loops during S phase to avert telomere catastrophe[7,8,14], while also preventing promiscuous t-loop unwinding during other cell cycle stages. We suggest that such

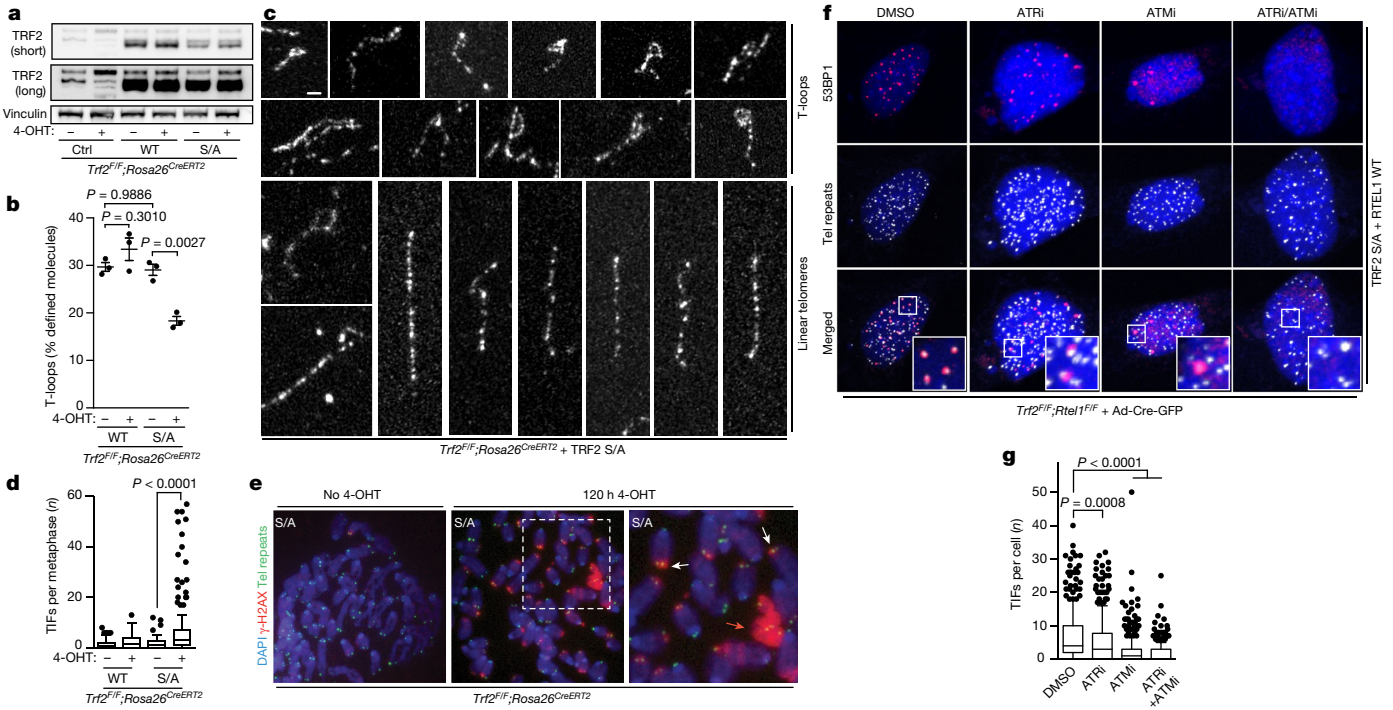


Fig. 4 | Expression of the TRF2(S367A) mutant promotes telomere-dysfunction-induced foci and impairs the formation of t-loops. a, Western blotting analysis of cells of the indicated genotype. Short and long indicate short and long exposure times. **b**, Quantification of t-loops observed in *Trf2*$^{F/F}$ MEFs expressing wild-type Myc–TRF2 or the Ser367Ala mutant with or without treatment with 4-hydroxytamoxifen (4-OHT) for 120 h. Data are exclusive of ambiguous molecules ($n = 3$ biological replicates scoring ≥1,192 molecules per replicate). Data are mean ± s.e.m. **c**, Representative images of t-loops and linear telomeres identified by Airyscan super-resolution imaging. Scale bar, 1 µm. **d**, Quantification of telomere-dysfunction-induced foci (TIF) per metaphase ($n = 120$ metaphases). Data are mean ± s.e.m. **e**, Representative images of TIF metaphase assays. TIF examples (white arrows) and mitotic catastrophe (red arrow) are shown. **f**, Representative images of TIF interphase assay with or without ATR or ATM inhibitors. Co-localization of 53BP1 and telomeres is shown in the merged panels. **g**, Quantification of TIF interphase assay in cells of the indicated genotype with or without ATR or ATM inhibitors (ATRi or ATMi, respectively) ($n > 300$ nuclei). Data are mean ± s.d. In **a** and **c**–**g**, the experiments were independently repeated at least three times with similar results. *P* values determined by one-way ANOVA (**b**, **d**, **g**). Box plots are as in Fig. 3e.

exquisite control of TRF2 to regulate t-loop opening and the need to 'protect' t-loops from promiscuous unwinding by RTEL1 outside of S phase, further demonstrate that t-loops are essential for physiological telomere homeostasis and chromosome end protection.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1744-8.

1. Doksani, Y., Wu, J. Y., de Lange, T. & Zhuang, X. Super-resolution fluorescence imaging of telomeres reveals TRF2-dependent T-loop formation. *Cell* **155**, 345–356 (2013).
2. Griffith, J. D. et al. Mammalian telomeres end in a large duplex loop. *Cell* **97**, 503–514 (1999).
3. Celli, G. B. & de Lange, T. DNA processing is not required for ATM-mediated telomere damage response after TRF2 deletion. *Nat. Cell Biol.* **7**, 712–718 (2005).
4. Benarroch-Popivker, D. et al. TRF2-mediated control of telomere DNA topology as a mechanism for chromosome-end protection. *Mol. Cell* **61**, 274–286 (2016).
5. Sfeir, A. et al. Mammalian telomeres resemble fragile sites and require TRF1 for efficient replication. *Cell* **138**, 90–103 (2009).
6. Cesare, A. J. & Griffith, J. D. Telomeric DNA in ALT cells is characterized by free telomeric circles and heterogeneous t-loops. *Mol. Cell. Biol.* **24**, 9948–9957 (2004).
7. Vannier, J. B., Pavicic-Kaltenbrunner, V., Petalcorin, M. I., Ding, H. & Boulton, S. J. RTEL1 dismantles T loops and counteracts telomeric G4-DNA to maintain telomere integrity. *Cell* **149**, 795–806 (2012).
8. Sarek, G., Vannier, J. B., Panier, S., Petrini, J. H. J. & Boulton, S. J. TRF2 recruits RTEL1 to telomeres in S phase to promote t-loop unwinding. *Mol. Cell* **57**, 622–635 (2015).
9. Chi, Y. et al. Identification of CDK2 substrates in human cell lysates. *Genome Biol.* **9**, R149 (2008).
10. Picco, V. et al. ERK1/2/MAPK pathway-dependent regulation of the telomeric factor TRF2. *Oncotarget* **7**, 46615–46627 (2016).
11. Vannier, J. B. et al. RTEL1 is a replisome-associated helicase that promotes telomere and genome-wide replication. *Science* **342**, 239–242 (2013).
12. Mendez-Bermudez, A. et al. Genome-wide control of heterochromatin replication by the telomere capping protein TRF2. *Mol. Cell* **70**, 449–461 (2018).
13. Van Ly, D. et al. Telomere loop dynamics in chromosome end protection. *Mol. Cell* **71**, 510–525 (2018).
14. Margalef, P. et al. Stabilization of reversed replication forks by telomerase drives telomere catastrophe. *Cell* **172**, 439–453 (2018).

Methods

Cell culture procedures

SV40-LT-immortalized *Rtel1*$^{F/F}$ MEFs[7], *Trf2*$^{F/-}$ MEFs (a gift from T. de Lange)[3], *Trf2*$^{F/F}$;*Rosa26*^{CreERT2} MEFs a gift from E. Lazzerini-Denchi)[15] and *Trf2*$^{F/F}$;*Rtel1*$^{F/F}$ mouse ear fibroblasts were cultured in DMEM supplemented with 15% fetal bovine serum (FBS; Invitrogen), L-glutamine and penicillin–streptomycin. HEK 293 and Phoenix Ampho 293 cells were kept in DMEM with 10% FBS. *Trf2*$^{F/F}$;*Rtel1*$^{F/F}$ conditional double-knockout mouse ear fibroblasts were isolated from adult mice obtained by crossing of the individual targeted *Trf2*$^{F/F}$ and *Rtel1*$^{F/F}$ mice. Genotypes were determined by Transnetyx using quantitative PCR with allele-specific probes. Production of retroviral supernatants and transductions of *Rtel1*$^{F/F}$ and *Trf2*$^{F/-}$ MEFs were done essentially as previously described[8]. *Trf2*$^{F/-}$ MEFs were infected with pLPC-puromycin retroviruses expressing control vector, Myc-tagged wild-type TRF2 or S367 mutants. *Trf2*$^{F/F}$;*Rosa26*^{CreERT2} MEFs were infected with pLPC-puromycin retroviruses expressing Myc-tagged wild-type TRF2 or mutant TRF2(S367A). Human HEK 293 cells were transduced with pLPC-puromycin retroviruses carrying Myc-tagged wild-type or mutant (S365A) TRF2. *Trf2*$^{F/F}$;*Rtel1*$^{F/F}$ mouse ear fibroblasts were complemented with pLPC-hygromycin retroviruses carrying Flag-tagged TRF2(S367A) and pBABE-puromycin retroviruses expressing mouse V5-tagged wild-type RTEL1 or RTEL1(R1237H) mutant. Transduced cells were selected with puromycin (2 μg ml^{-1}) for 2–6 days. *Trf2*$^{F/F}$;*Rtel1*$^{F/F}$ mouse ear fibroblasts were kept under puromycin (2 μg ml^{-1}) and hygromycin (150 μg ml^{-1}) selection for 5 days. Deletion of floxed alleles in *Rtel1*$^{F/F}$ and *Trf2*$^{F/-}$ MEFs was carried out with Ad-Cre-GFP adenovirus (Vector Biolabs) and cells were genotyped by PCR at 96 h after infection as previously described[3,7]. TRF2 was deleted in *Trf2*$^{F/F}$;*Rosa26*^{CreERT2} MEFs by adding 1 μM 4-hydroxytamoxifen (Sigma-Aldrich) to the culture medium. Cell lines were routinely tested for mycoplasma contamination with negative results.

Cell lysis, western blotting, immunoprecipitation and drug treatments

Cells were rinsed twice with PBS, transferred to an ice-cold NET lysis buffer (50 mM Tris (pH 7.2), 150 mM NaCl, 0.5% NP-40, 1× EDTA-free Complete protease inhibitor cocktail (Roche), 1× PhosSTOP phosphatase inhibitor cocktail (Roche)) and lysed for 10 min on ice. The cell lysates were then briefly vortexed and passed through a 23G syringe five times. The soluble protein fractions were collected after centrifugation at 16,000*g* for 10 min at 4 °C. Western blotting analysis was performed as previously described[7]. Immunoblots of whole-cell extracts from *Trf2*$^{F/F}$;*Rosa26*^{CreERT2} cells with or without exogenous expression of the Myc–TRF2 allele was performed as previously described[13]. For protein immunoprecipitation, whole-cell extracts were precleared with protein G Sepharose (Sigma-Aldrich) and 1–2 mg of precleared extract was incubated with the indicated antibodies. Immunocomplexes were subjected to SDS–PAGE followed by immunoblotting using nitrocellulose membrane (GE Healthcare). See Supplementary Table 1 for a list of antibodies used. For inhibition of CDKs, R-roscovitine (Sigma-Aldrich), was used at a final concentration of 10 μM for 24 h. PLK-1 was inhibited by BI-2536 (Axon Medchem) at a final concentration of 100 nM for 24 h. The MEK–ERK signalling pathway was inhibited by the MEK1 and MEK2 inhibitor U0126 (Selleckchem), at a concentration of 30 μM for 24 h. An equal amount of DMSO was used as a vehicle control.

siRNA treatment and siRNA oligonucleotides

Transfections with siRNA oligonucleotides were performed using the Lipofectamin RNAiMax (Thermo Fisher Scientific). In brief, human cells at density of 2.0×10^5 cells per well were transfected in a 6-well plate with 40 pmol siRNA. Mouse cells at density of 3.0×10^5 cells per well were transfected with 150 pmol siRNA. The medium was exchanged 24 h after transfection. Then, 72 h after transfection, the cells were collected and the levels of proteins of interest were assessed by immunoblot analyses as described. For silencing experiments in human and mouse cells, pre-designed SMARTpool ON-TARGETplus and Accell siRNA oligonucleotides (Dharmacon; GE Healthcare) were used, respectively. For siRNA oligonucleotides details, see Supplementary Table 2.

λ-Phosphatase treatment

Whole-cell extracts were prepared as described above except in the absence of phosphatase inhibitors. Lysates were incubated with 800 U of λ-phosphatase (New England Biolabs) in NET lysis buffer supplemented with 1 mM of $MnCl_2$ along with protease inhibitors for 30 min at 30 °C. Next, the lysates were incubated on ice for 15 min and subjected to immunoprecipitation as detailed in the main text.

In vitro kinase assay

Whole-cell extracts from HEK 293 cells were incubated for 1 h at 4 °C with the rabbit polyclonal anti-RTEL1 antibody. Immunocomplexes were coupled to protein G Sepharose beads for an additional 1 h at 4 °C and washed three times with the NET lysis buffer followed by two washes with kinase buffer (20 mM Tris, pH 7.5, 50 mM KCl, 7.5 mM $MgCl_2$, 10 mM $MnCl_2$, 1 mM DTT, 1× PhosSTOP phosphatase inhibitors). Kinase reactions were performed by incubating the immunocomplexes with 20 μl of kinase buffer containing cold adenosine triphosphate (ATP) and 1 μg recombinant CDK2–cyclinA protein complex for 20 min at 37 °C. Reactions were washed twice with kinase buffer and terminated by the addition of 5× SDS–PAGE sample buffer, and resolved by SDS–PAGE.

Generation of pTRF2(Ser365/367) phospho-specific antibodies

Rabbit polyclonal antibodies against human TRF2 phosphorylated at Ser365 and mouse TRF2 phosphorylated at Ser367 were generated by Kaneka Eurogentec S.A. Biologics Division. The antibodies were raised against the phosphorylated human C-(PTQALPA[pS]PALKNKR)-N and mouse C-(ANLASPS[pS]PAHKHKR)-N TRF2 sequences conjugated through the added C-terminal cysteine to keyhole limpet hemocyanin (KLH). Phosphoserine 365- and 367-specific antibodies were purified with the use of the corresponding sulfolinked phospho- and unphosphorylated peptides. The specificity of each antibody was confirmed by ELISA and immunoblot assays.

Peptide synthesis and peptide pull-down experiments

The peptide pull-down was carried out using the biotinylated peptides. In brief, 36 μg of each of the peptides was coupled to 40 μl of streptavidin-coated magnetic beads (Invitrogen) and added to 1 mg of nuclear extract of HEK 293 cells expressing pHAGE-HA-Flag-RTEL1. Nuclear extracts were precleared by incubation for 30 min at room temperature with uncoupled beads before pull-down incubation. The coupled beads and the lysates were incubated for 2 h at 4 °C. The beads were washed four times with TBST (Tris-buffered saline, 0.1% Tween-20), resuspended in 2× SDS loading sample buffer, and boiled for 5 min.

Slot-blot assay

TRF2 peptides diluted into a final volume of 200 μl in SSC 2× were applied under gentle vacuum to Trans-Blot nitrocellulose membrane (Bio-Rad) using a Minifold 48 slots, Whatman apparatus (GE Healthcare). Each well was washed with 200 μl aliquots of SSC 2×. After removing SSC 2× with gentle suction, the membrane was removed from the apparatus and washed once with SSC 2×. The membrane was blocked at room temperature in a blocking buffer (5% BSA in TBST) for 1 h and probed with horseradish peroxidase (HRP)-conjugated anti-biotin antibody. Incubation was allowed to proceed for 1 h at 4 °C with rocking. After incubation, the antibody solution was removed and the membrane rinsed twice with TBST followed by detection by the ECL method.

Site-directed mutagenesis

Amino acid substitutions were performed with the primers as indicated in key resources table. Primers were designed with the QuikChange

Primer Design Software (Agilent Technologies). Single mutants were generated using the QuikChange Lightning Site-Directed Mutagenesis kit (Agilent Technologies) and double mutants were created with the QuikChange Lightning Multi Site-Directed Mutagenesis kit (Agilent Technologies) according to the manufacturer's instructions. The generated mutants were verified by sequencing to screen against spurious secondary mutations. For primer sequences, see Supplementary Table 2.

In situ PLA

Cells were plated on coverslips at density 5×10^4 in 24-well plates and left in culture conditions overnight. The next day, cells were pre-extracted in CSK buffer (10 mM PIPES, pH 6.8, 100 mM NaCl, 300 mM sucrose, 3 mM $MgCl_2$, 1 mM EGTA and 0.5% Triton X-100) fixed with 4% formaldehyde in the CSK buffer for 10 min, permeabilized with PBS containing 0.5% (v/v) NP-40 for 5 min, and blocked 30 min with goat serum (5%) in PBS. PLA was performed following the manufacturer's instructions using the Duolink anti-Mouse MINUS and anti-Rabbit PLUS In Situ PLA probes and the Duolink In Situ Detection Reagents Red (Olink Bioscience). Images were acquired with a Zeiss Axio Imager M1 microscope equipped with an ORCA-ER camera (Hamamatsu) and using the Volocity 6.3 software (Perkin Elmer).

Cell cycle synchronization

HEK 293 cells were synchronized by the double-thymidine-block method with minor modifications. In brief, cells were treated with 2 mM thymidine (Sigma-Aldrich) for 18 h, thymidine-free medium for 9 h to release the cells, and 2 mM thymidine was added to medium for an additional 16 h to arrest the cells at the G1-to-S transition. Cells were washed twice with PBS and then released in fresh complete DMEM. Cells were analysed at 70-min time intervals by immunoblotting and in situ PLA assay. For synchronization in mitosis, a thymidine–nocodazole block was used. In brief, cells at a confluence of 60% were treated with 2 mM thymidine for 24 h, washed twice in PBS, and released into complete DMEM for 3 h. Next, cells were treated with 50 μg ml^{-1} of nocodazole (Sigma-Aldrich) for 15 h, and the cells were washed twice with PBS and a fresh complete medium was added to the cell culture. Synchronized cells were analysed at 90-min time intervals by western blotting with antibodies as indicated.

Indirect immunofluorescence

Cells were washed with PBS and fixed with 4% formaldehyde for 10 min at room temperature, permeabilized with 0.3% Triton X-100 in PBS for 5 min at room temperature and then blocked with 3% BSA, 10% FBS in PBS for 1 h at room temperature. Samples were then incubated with rabbit anti-53BP1 antibody overnight at 4 °C, washed with 0.05% Tween-20 in PBS and incubated with anti-rabbit IgG AlexaFluor 594 (Molecular Probes). DNA was counterstained with DAPI and images were acquired using a Zeiss AxioImager M1, using a Hamamatsu digital camera and the Volocity 4.3.2 software (Perkin Elmer).

Airyscan super-resolution imaging

Sample preparation for super-resolution microscopy, cross-linking efficiency determination and Airyscan imaging were performed as described previously[13].

Super-resolution microscopy scoring criteria

Images obtained from super-resolution microscopy were scored as previously described[1,13]. Specifically, after capture and processing, images were exported to ImageJ as .tif images with maintained scales. Images were manually quantified with researchers blinded to the experimental conditions. Telomere molecules were scored if they had a traceable telomere contour of ≥1 μm, and contained no gaps in telomere staining ≥500 nm. Molecules were classified as t-loops when we could discern an individual molecule consisting of a closed loop structure with a single attached tail. Molecules were classified as linear when we observed an individual molecule with two visible ends, containing no loops or branched structures. All molecules that did not conform to the looped or linear definition were classified as ambiguous. Densely packed areas of coverslips with overlapping telomere molecules were not scored. Each loop and linear molecule were measured for contour length using the ImageJ trace function.

PNA FISH and inmmuofluorescence FISH

Telomeric PNA FISH on cytogenetic chromosome spreads was performed as previously described[7]. In brief, cells were treated with 0.2 μg ml^{-1} of colcemid for 90 min to arrest cells in metaphase. Trypsinized cells were incubated in 75 mM KCl, fixed with methanol:acetic acid (3:1), and spread on a glass slide. To preserve the chromosome architecture better, the slides were rehydrated in PBS for 5 min, fixed in 4% formaldehyde for 5 min, treated with 1 mg ml^{-1} pepsin for 10 min at 37 °C, and fixed in 4% formaldehyde for 5 min. Next, slides were dehydrated in 70%, 85% and 100% (v/v) ethanol for 15 min each and air-dried. Metaphase chromosome spreads were hybridized with telomeric TAMRA-TelG 5′-$(TTAGGG)_3$-3′ PNA probe (Panagene) and slides were mounted using ProLong Gold antifade with DAPI (Life Technologies). Chromosome images and telomere signals were captured using Zeiss Axio Imager M1 microscope equipped with an ORCA-ER camera (Hamamatsu) controlled by Volocity 6.3 software (Improvision). For interphase immunofluorescence FISH (TIFs), cells grown on #1.5 glass coverslips were fixed for 20 min in 2% (w/v) formaldehyde (Thermo Scientific) at room temperature and immunofluorescence FISH was performed as previously described[7], using primary 53BP1 antibody (see Supplementary Table 1), anti-mouse Alexa Fluor secondary antibody (Molecular Probes) and a TAMRA-TelG 5′-$(TTAGGG)_3$-3′ PNA probe (Panagene). Metaphase TIF assays were done as previously described[16]. In brief, cells were treated with 20 ng ml^{-1} colcemid for 1 h before collecting and resuspending in a hypotonic buffer of 0.2% trisodium citrate in 0.2% KCl. The cells were swollen for 5 min then cytocentrifuged onto glass slides using a Tharmac Cellspin 1, before fixation and processed for immunolabelling with an anti-γ-H2AX primary antibody and subsequently with anti-mouse Alexa Fluor 568 secondary antibody (Molecular Probes). After a second fixation, the samples were hybridized with an Alexa Fluor 488 conjugated C-rich telomere PNA probe (Panagene), stained with DAPI and mounted with ProLong Gold (Molecular Probes). Automated metaphase finding and image capture was done as previously described[17] using a MetaSystems imaging platform coupled with a ZEISS AxioImager Z.2 microscope using a 63×, 1.4 NA oil objective and appropriate filter cubes, and a CoolCube1 camera (MetaSystems). After acquisition, images were imported into ImageJ (NIH) and Adobe Photoshop CS5 for manual quantification and processing.

Telomere circle assay

Cells grown at a confluence between 70% and 80% were collected from two 10-cm dishes and extraction of genomic DNA for T-circle assay was performed as previously described[7]. Total gDNA was digested by AluI/HinfI restriction enzymes and the TCA assay was performed with two essential modifications as described[7]: (1) Phi29 DNA (Thermo Scientific) polymerization used a mammalian telomere primer; and (2) Southern blotting membrane was hybridized to a γ^{32}P-labelled $(TTAGGG)_4$ telomeric probe. Southern blot images were captured with Storm 840 scanner and the extent of [^{32}P] incorporation was quantified from the autoradiographs by ImageQuant TL Software Analyzer (Amersham Biosciences). The level of γ[^{32}P] incorporation obtained from the Phi29-negative control samples represented the background level, which was subtracted from the values obtained from the samples that contained the Phi29 DNA polymerase.

DNA combing

Cells were sequentially pulse-labelled with 25 μM CldU (Sigma) and 250 μM IdU (Sigma) for 20 min and, after collection of the cells, low-melting agarose (Sigma) plugs each containing 200,000 cells were prepared. DNA fibres were extracted from the plugs and combed on silanized

coverslips using the FiberPrep DNA extraction kit and the molecular combing system (Genomic Vision), according to the manufacturer's instructions. Combed fibres were fixed at 60 °C for 2 h and DNA was denaturated in 0.5 M NaOH for 25 min. Fibres were then blocked in 1% BSA in 0.1% Tween-20 in PBS for 1 h, incubated with rat anti-BrdU (detects CldU, BU1/75, AbD Serotec, 1:200) and mouse anti-BrdU (detects IdU, B44, BD Biosciences, 1:200) for 1 h followed by anti-rat IgG AlexaFluor 594 and anti-mouse IgG AlexaFluor 488 (Molecular Probes, 1:500) for 1.5 h.

For DNA combing with single-stranded DNA (ssDNA) staining, cells were sequentially pulse-labelled with 25 μM CldU (Sigma) and 250 μM IdU (Sigma) for 15 min, and, after collection of the cells, low-melting agarose (Sigma) plugs each containing 250,000 cells were prepared. DNA fibres were extracted from the plugs and combed on silanized coverslips using the FiberPrep DNA extraction kit and the molecular combing system (Genomic Vision), according to the manufacturer's instructions. Combed fibres were fixed at 60 °C for 2 h and DNA was denaturated in 0.5 M NaOH for 25 min. Fibres were then blocked in 1% BSA in 0.1% Tween-20 in PBS for 1 h, incubated with rat anti-BrdU (detects CldU, abcam, ab6326, 1:500) and mouse anti-BrdU (detects IdU, B44, BD Biosciences, 1:250) for 1 h followed by anti-rat IgG AlexaFluor 594 and anti-mouse IgG AlexaFluor 488 (Molecular Probes, 1:500) for 1.5 h. Fibres were then incubated with mouse anti-ssDNA antibody (Millipore, MAB3034, 1:200) for 45 min followed by anti-mouse IgG AlexaFluor 647 (Molecular Probes, 1:200) for 45 min.

Images were acquired using a Zeiss AxioImager M1, equipped with Hamamatsu digital camera and the Volocity software (Perkin Elmer). Fibre length was analysed using ImageJ (http://rsbweb.nih.gov/ij/).

Mass spectrometric analyses and protein identification

Coomassie-stained polyacrylamide gel slices were excised from SDS–PAGE gels using a scalpel and processed for mass spectrometry using the Janus liquid handling system (PerkinElmer). In brief, the excised protein gel pieces were placed in individual wells of a 96-well microtitre plate and destained with 50% (v/v) acetonitrile and 50 mM ammonium bicarbonate, reduced with 10 mM DTT, and alkylated with 55 mM iodoacetamide. After alkylation, the samples were digested with trypsin (Promega), overnight at 37 °C. The resulting peptides were extracted in 1% (v/v) formic acid, 2% (v/v) acetonitrile. Digests were subsequently analysed by nano-scale capillary LC–MS/MS. Peptide mixtures were separated on a 50 cm, 75 μm i.d. EasySpray C18 LC-MS column over a 30-min gradient and eluted directly into the LTQ Orbitrap Velos (Thermo Scientific) mass spectrometer. The mass spectrometer was operated in data dependent mode with the top-10 most-intense multiply charged precursor ions fragmented in the linear ion trap using collision-induced dissociation. Raw mass spectrometric data were processed in MaxQuant[18] (v.1.3.0.5) for peptide and protein identification, the database search was performed using the Andromeda search engine against the *Homo sapiens* canonical sequences downloaded from UniProtKB (release 2012_08).

Statistical analysis

Statistical analyses were performed using GraphPad PRISM version 7.0 software (GraphPad). Statistical significance of data was assessed by two-tailed Student *t*-test or one-way ANOVA unless noted otherwise. Data represent mean ± s.e.m. or mean ± s.d. as indicated. $P > 0.05$ was considered not significant. No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment unless stated otherwise.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The mass spectrometry proteomics dataset is publicly available through ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD014843. Source Data for Figs. 1–4 and Extended Data Figs. 1–8 are available with the online version of the paper. All other data are available from the corresponding author upon reasonable request.

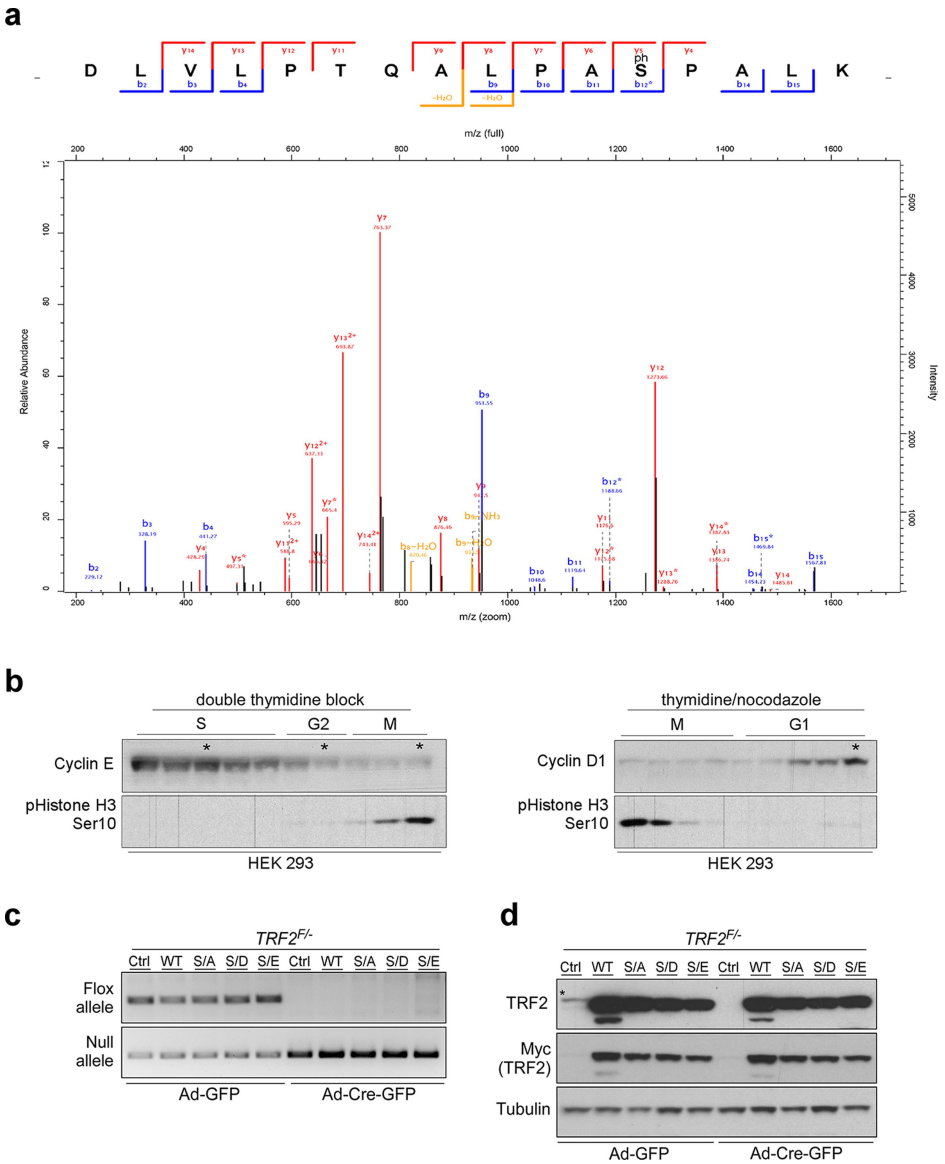
15. Okamoto, K. et al. A two-step mechanism for TRF2-mediated chromosome-end protection. *Nature* **494**, 502–505 (2013).
16. Cesare, A. J., Heaphy, C. M. & O'Sullivan, R. J. Visualization of telomere integrity and function in vitro and in vivo using immunofluorescence techniques. *Curr. Protoc. Cytom.* **73**, 12.40.1–12.40.31 (2015).
17. Cesare, A. J., Hayashi, M. T., Crabbe, L. & Karlseder, J. The telomere deprotection response is functionally distinct from the genomic DNA damage response. *Mol. Cell* **51**, 141–155 (2013).
18. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).

Acknowledgements We thank members of the Boulton and Cesare laboratories for suggestions, discussions and critical reading of the manuscript. We thank N. O'Reilly and D. Joshi for peptide synthesis, the Australian Cancer Research Foundation Telomere Analysis Centre at the Children's Medical Research Institute (Sydney) for imaging support and A. Colomba for providing reagents. G.S. is supported by an EMBO advanced fellowship (ALTF 1656-2014). P.K. and P.R. are supported by the Crick Institute core funding. The work in the Chowdhury laboratory is supported by the National Institutes of Health (NIH) R01 CA208244. The work in the Boulton laboratory is supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC0010048), the UK Medical Research Council (FC0010048), and the Wellcome Trust (FC0010048); a European Research Council (ERC) Advanced Investigator Grant (TelMetab); and a Wellcome Trust Senior Investigator Grant. The Cesare laboratory is supported by National Health and Medical Research Council of Australia (1106241) and the Cancer Institute NSW (11/FRL/5-02).

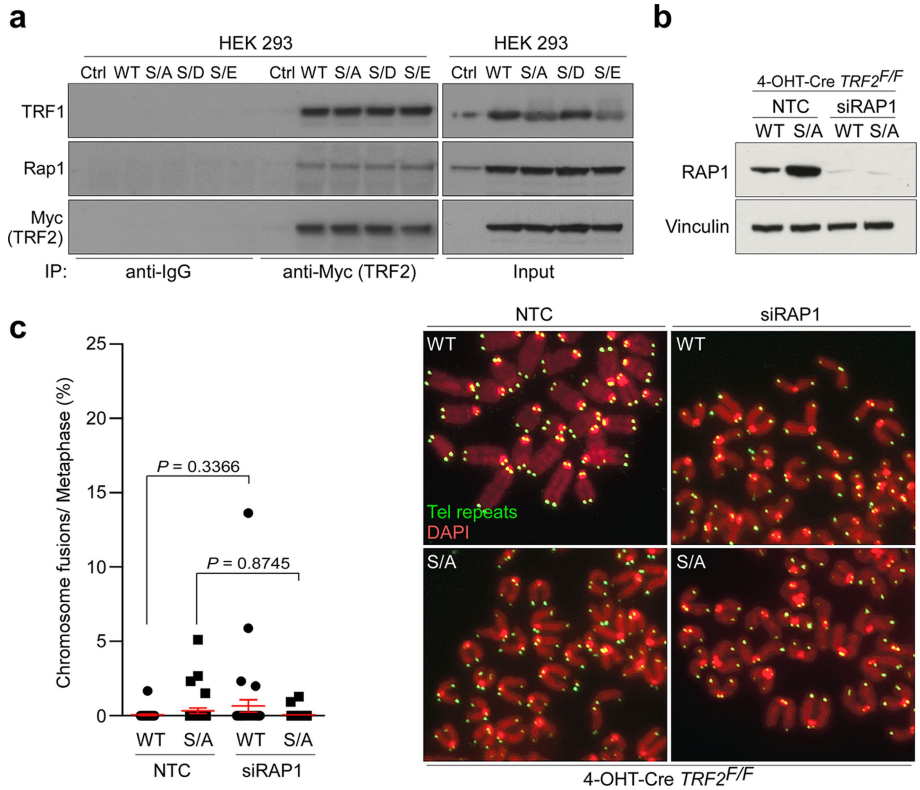
Author contributions S.J.B. and G.S. conceived the study; G.S., P.K., P.R., D.V.L., A.J.C. and S.J.B. designed experiments; G.S., P.K., P.R., D.V.L., P.M., V.B., X.-F.Z. and H.R.F. conducted experiments; G.S., P.K., P.R., D.V.L., P.M., X.-F.Z., H.R.F., A.P.S., D.C., A.J.C. and S.J.B. analysed data; S.J.B., G.S. and A.J.C. wrote the manuscript with editorial help from P.K., P.R., P.M. and D.V.L.

Competing interests The authors declare no competing interests.

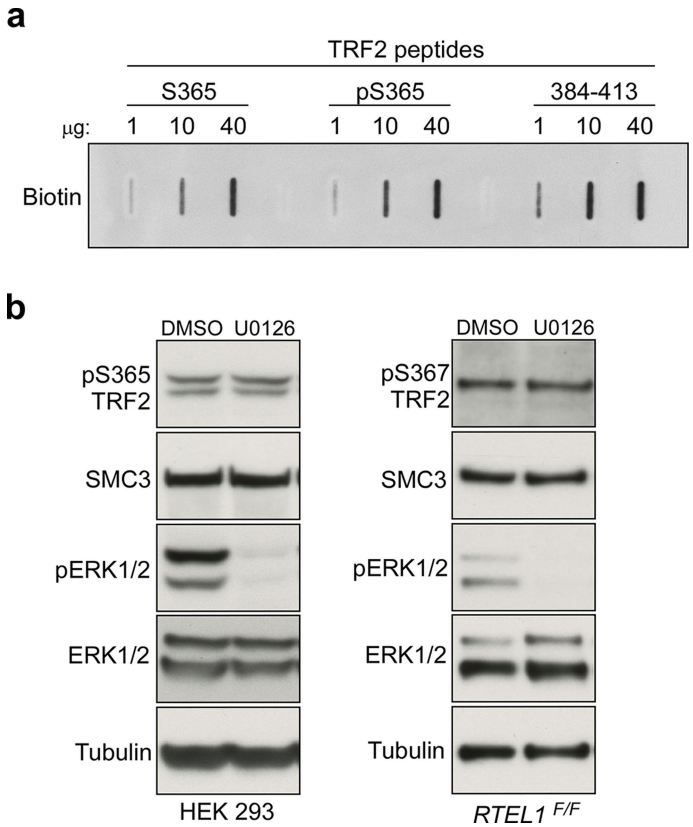
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1744-8.
Correspondence and requests for materials should be addressed to S.J.B.
Peer review information *Nature* thanks Eric Gilson, Joachim Lingner and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



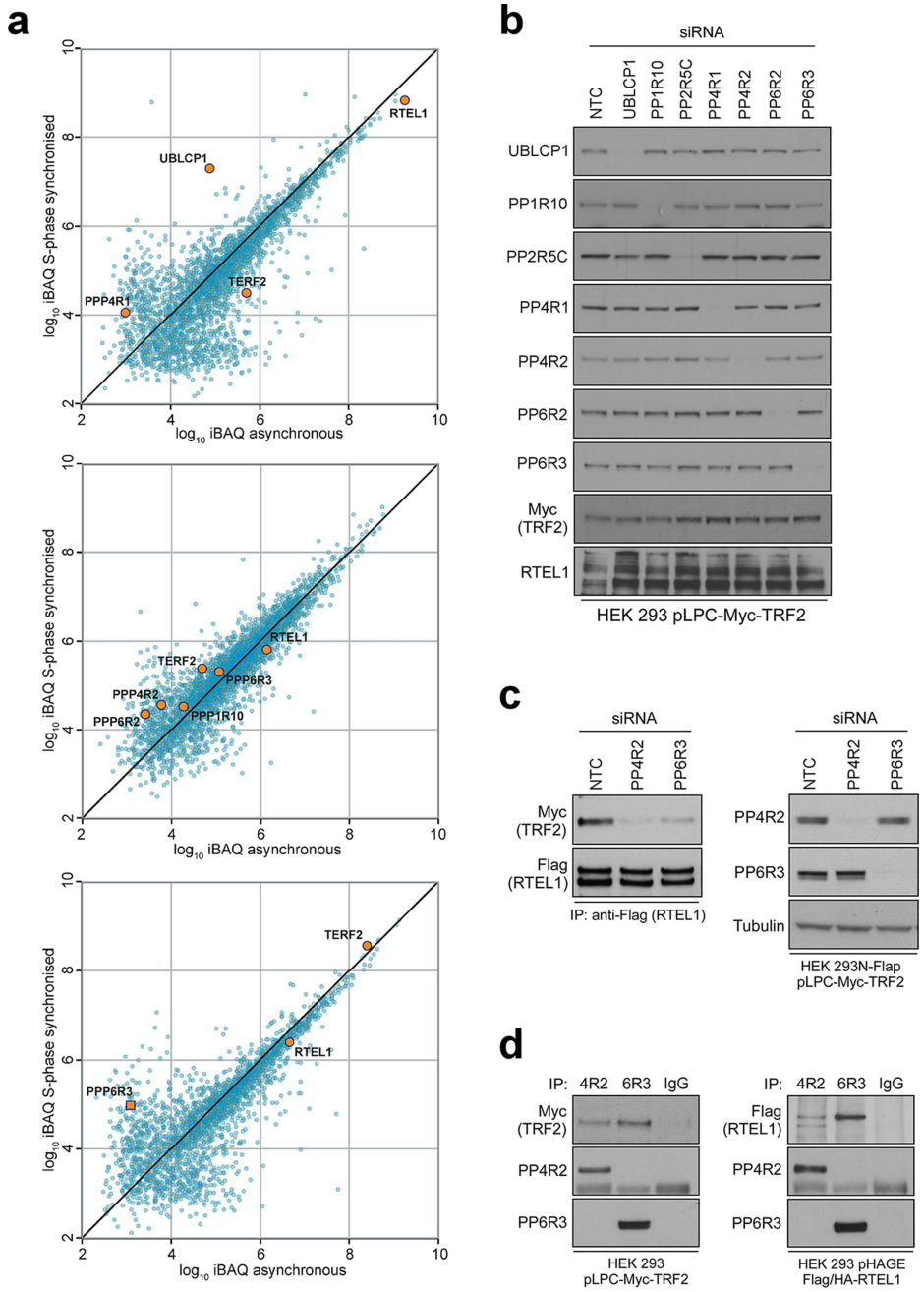
Extended Data Fig. 1 | TRF2 is phosphorylated at Ser365. a, Annotated spectrum for the TRF2 phosphorylated peptide. The data were acquired on the LTQ Orbitrap Velos and processed in MaxQuant v.1.3.0.5 with the database search performed against the canonical sequences *Homo sapiens* from UniProt. For the spectrum shown, the posterior error probability value was 0.018258 and the localization score for the site was 1 DLVLPTQALPAS(1)PALK. **b**, HEK 293 cells were released from a double-thymidine block (left) or a thymidine plus nocodazole block (right). Cells were subjected to SDS–PAGE analysis and progression through the cell cycle was monitored by immunoblotting with cell cycle markers as indicated. Asterisks indicate time points after synchronization. **c**, PCR analysis of genomic DNA isolated from *Trf2*$^{F/-}$ MEFs stably expressing empty vector, wild-type or mutant TRF2, 96 h after infection with control or Cre-expressing adenovirus. **d**, Western blotting analysis of the cells described in **c** to monitor loss of endogenous TRF2 after Cre expression and to determine complementation efficiency with ectopic wild-type and mutant TRF2. The asterisk indicates endogenous TRF2. In **a**–**d**, the experiments were independently repeated at least twice with similar results.



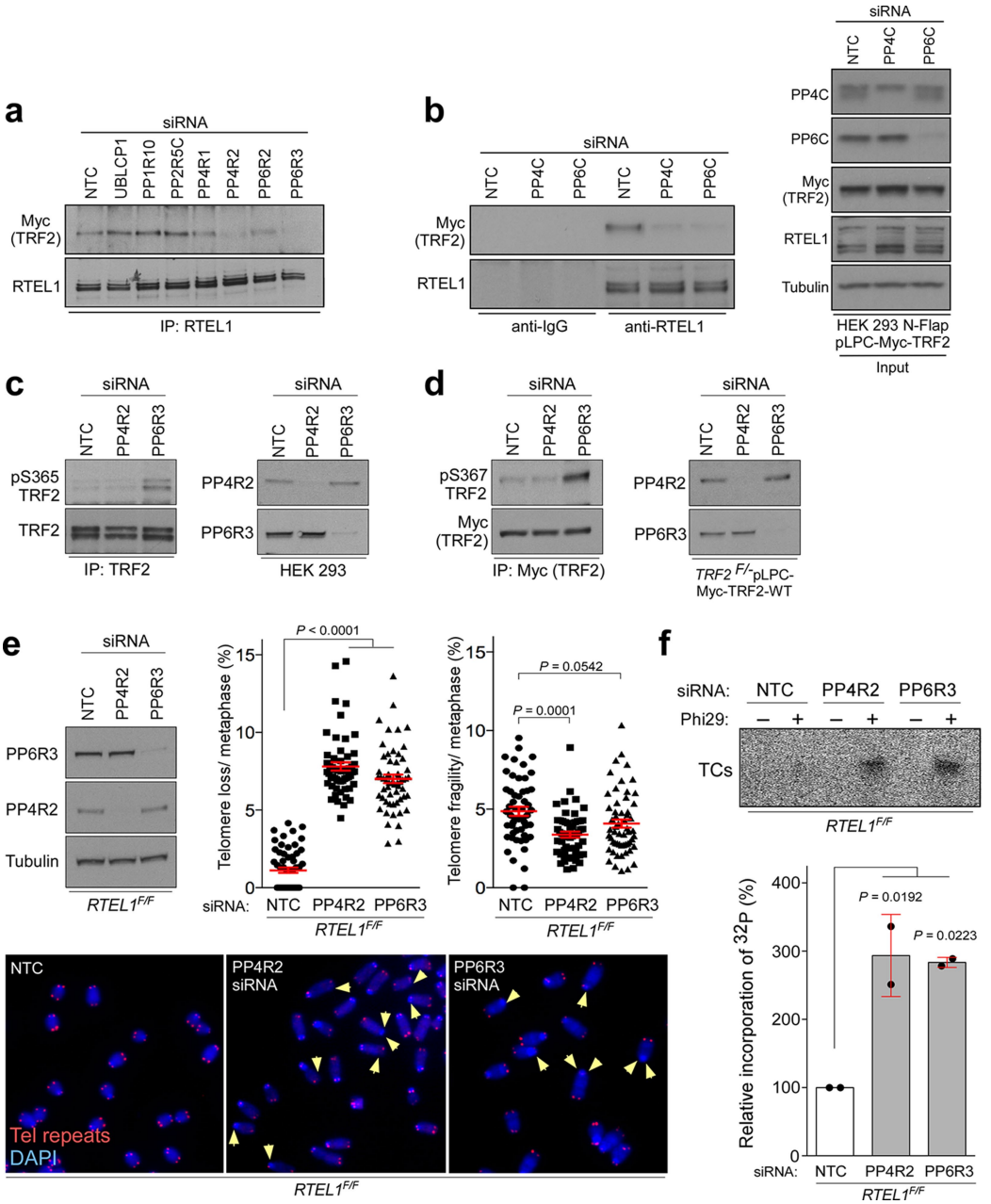
Extended Data Fig. 2 | Mutations of TRF2 at the Ser365 or Ser367 phospho-site do not affect interaction with shelterin components. a, Whole-cell extracts from HEK 293 cells stably expressing empty vector, wild-type or mutant Myc-tagged TRF2 as indicated were immunoprecipitated with anti-Myc antibody or normal mouse IgG. Protein complexes were analysed with antibodies against RAP1, TRF1 and Myc. **b**, *Trf2*$^{F/F}$ MEFs expressing wild-type or phospho-dead (Ser367Ala mutant) TRF2 were transfected with either control siRNA (non-targeting control, NTC) or siRNA against *RAP1* (siRAP1) and treated with 4-OHT for 96 h. Whole-cell extracts were analysed 72 h later as indicated. **c**, Quantification (left) and representative images (right) of chromosome fusions in the *Trf2*$^{F/F}$ MEFs depicted in **b** performed 96 h after 4-OHT treatment ($n = 30$ metaphases analysed). Data are mean ± s.e.m. *P* values were determined by one-way ANOVA. In **a**–**d**, the experiments were independently repeated at least twice with similar results.



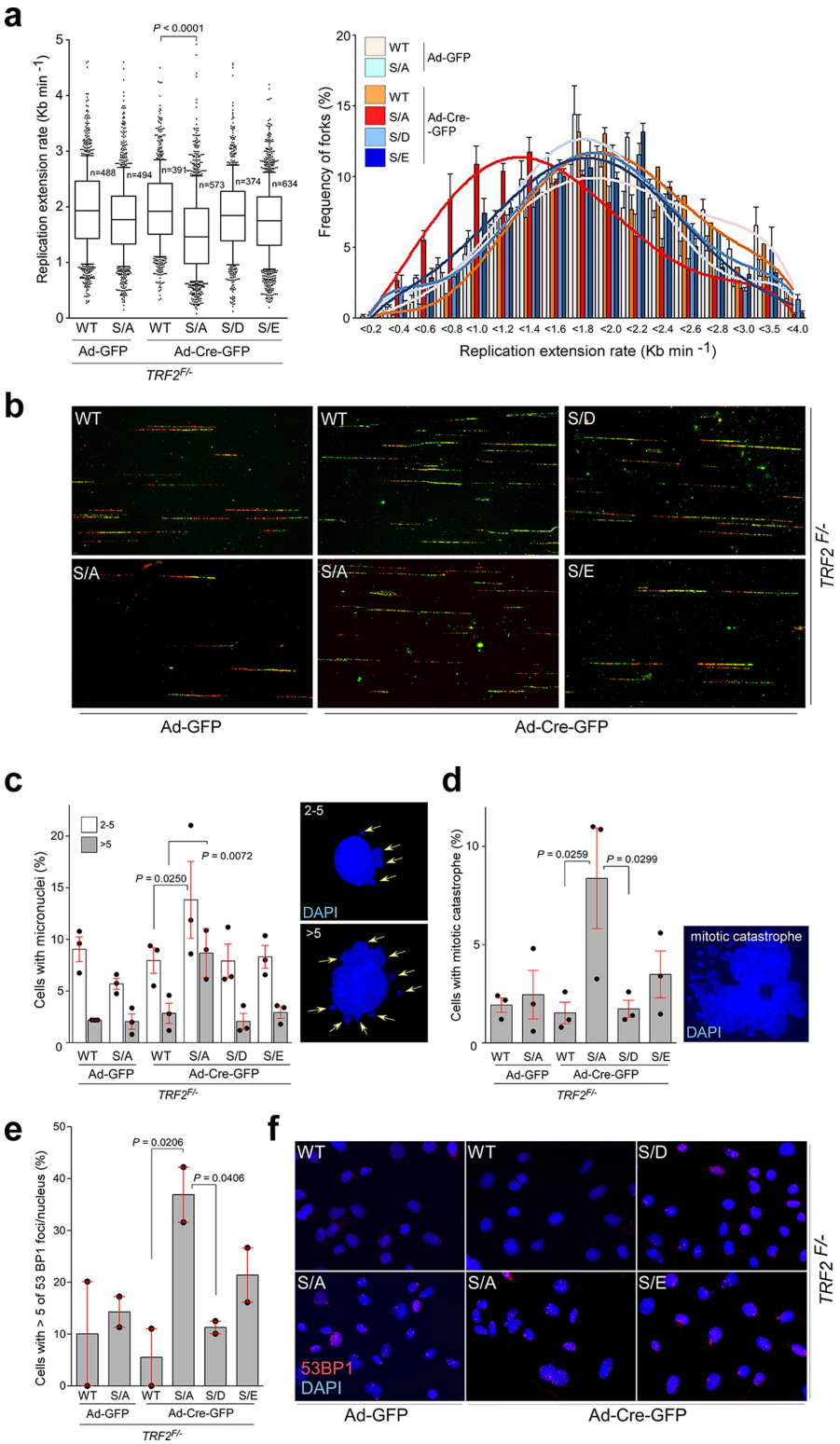
Extended Data Fig. 3 | Inhibition of MEK–ERK signalling pathway does not affect TRF2 phosphorylation at Ser365 or Ser367. a, Quantity screen for TRF2-biotinylated peptides. Slot-blot assay in which biotin-tagged TRF2 peptides were incubated with streptavidin-coated beads to ensure that the correct amounts were used in the peptide pull-down assay. **b**, HEK 293 cells (left) or *Rtel1*$^{F/F}$ MEFs (right) were pre-treated with vehicle control (DMSO) or with 25 μM of MEK1/2 kinase inhibitor (U0126) for 48 h. Whole-cell extracts were subjected to SDS–PAGE analysis followed by immunoblotting with antibodies as indicated. In **a** and **b**, the experiments were independently repeated at least twice with similar results.



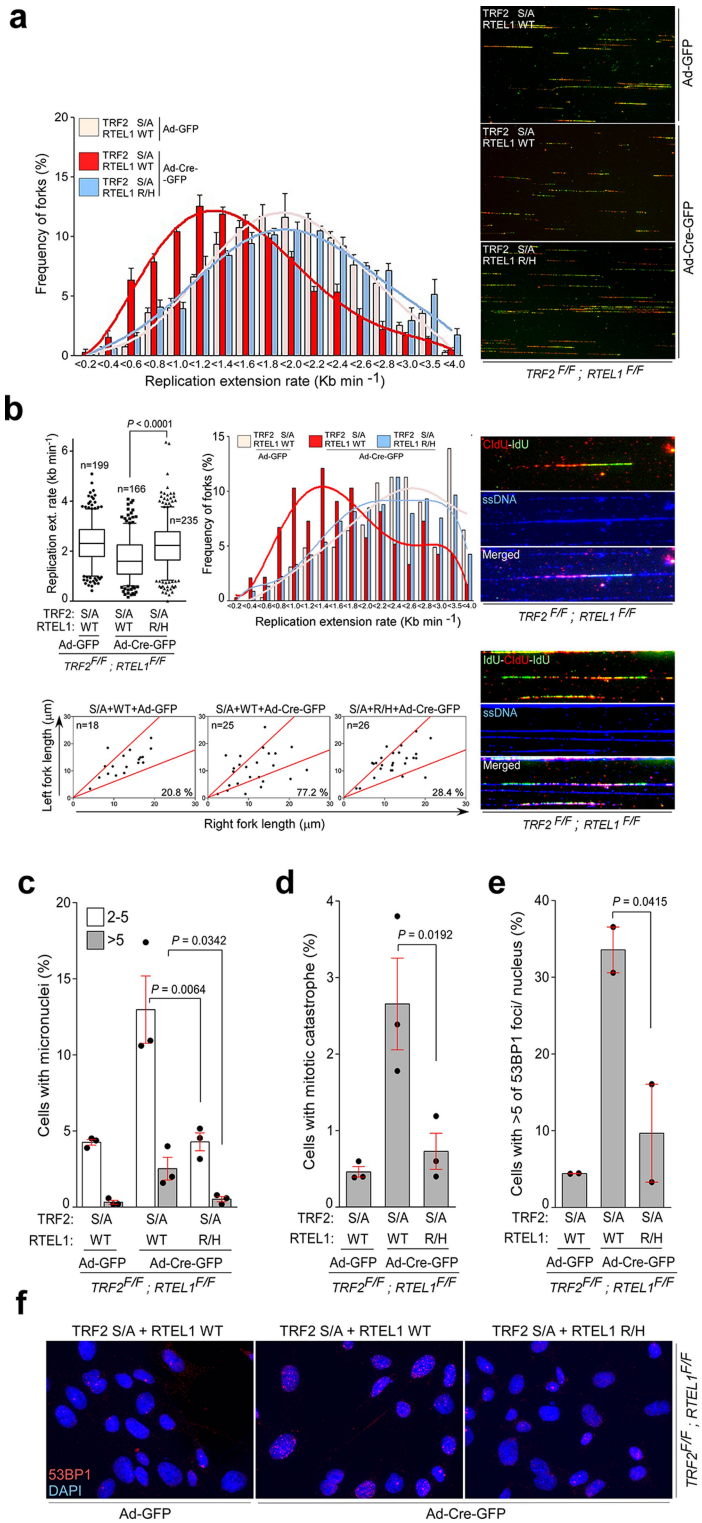
Extended Data Fig. 4 | Identification of TRF2- and RTEL1-interacting phosphatases and protein phosphatase regulatory subunits. a, Intensity-based absolute quantification (iBAQ) scatter plots comparing protein abundance in cells synchronized during S phase versus asynchronous control cells. Immunoprecipitates from asynchronous or S-phase-synchronized HEK 293 cells stably expressing Flag–haemagglutinin (HA)-tagged RTEL1 (top), N-terminal FLAP (Flag–GFP)-tagged RTEL1 (middle) or Myc–TRF2 (bottom) were separated by SDS–PAGE and stained with Coomassie blue to visualize proteins. Immunoprecipitations with haemagglutinin (top), GFP (middle) and Myc (bottom) antibodies were performed. The proteins along the entire length of the gel were extracted and analysed by liquid chromatography–tandem mass spectrometry (LC/MS–MS). **b**, HEK 293 cells stably expressing wild-type Myc–TRF2 were transfected with either non-target control or siRNA against protein phosphatase regulatory subunits, as specified. Three days later, protein levels were analysed with the indicated antibodies. **c**, FLAP-tagged RTEL1 HEK 293 cells expressing Myc-tagged wild-type TRF2 were transfected with either control siRNA or siRNA against *PP4R2* or *PP6R3*. Whole-cell extracts were immunoprecipitated with anti-Flag antibody and immunocomplexes were analysed for Myc (TRF2) and Flag (RTEL1). Inputs (5%) are shown on the right. **d**, HEK 293 cells expressing wild-type Myc–TRF2 (left) or Flag–HA-tagged RTEL1 (right) were subjected to immunoprecipitation with normal rabbit IgG or antibodies against PP4R2 and PP6R3. Immune complexes were analysed by western blotting with the indicated antibodies. In **b**–**d**, the experiments were independently repeated at least twice with similar results.



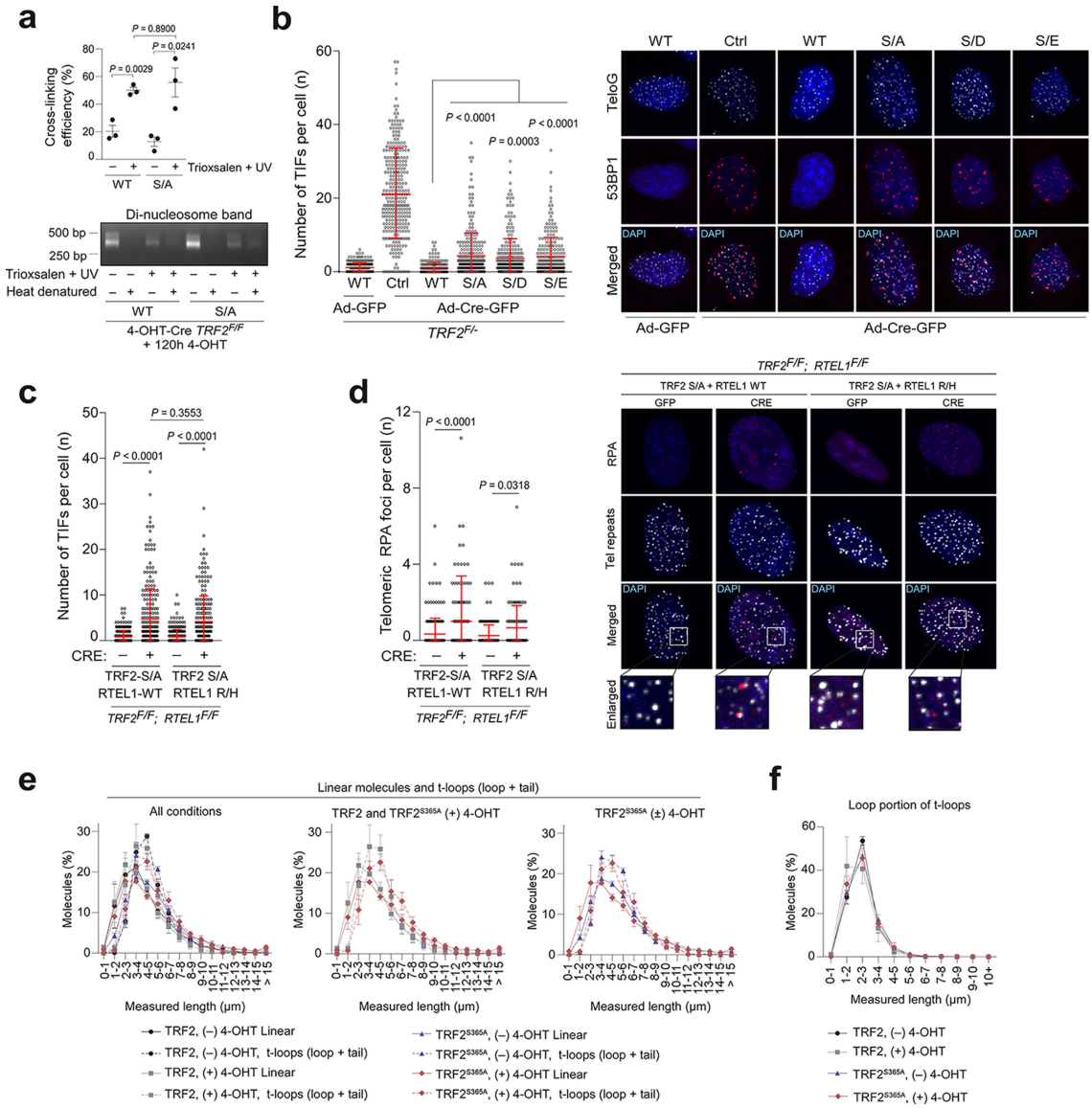
Extended Data Fig. 5 | PP6R3 controls phosphorylation of TRF2 at Ser365 or Ser367. HEK 293 cells expressing wild-type Myc–TRF2 were transfected with a non-targeting control siRNA or siRNAs against protein phosphatase regulatory subunits (**a**) or catalytic subunits (**b**). Cells were collected, and whole-cell extracts were immunoprecipitated with anti-RTEL1 antibody. Immunocomplexes were resolved by SDS–PAGE and analysed by western blotting as indicated. **c**, HEK 293 cells (**c**) and *Trf2*$^{F/-}$ MEFs (**d**) expressing Myc-tagged wild-type TRF2 were transfected with control siRNA or siRNA targeting *PP4R2* or *PP6R3* (*Pp4r2* or *Pp6r3* for MEFs). Whole-cell extracts were immunoprecipitated with anti-TRF2 antibody, and immunocomplexes were resolved by SDS–PAGE and analysed for human phospho-TRF2 (pS365 TRF2; left panel in **c**) or mouse phospho-TRF2 (pS367 TRF2; left panel in **d**). **e**, Top, frequency of telomere loss and telomere fragility per metaphase in *Rtel1*$^{F/F}$ MEFs transfected with control siRNA or with *Pp4r2* or *Pp6r3* siRNA ($n = 58$ (NTC), $n = 57$ (*Pp4r2*), and $n = 55$ (*Pp6r3*) of analysed metaphases). Efficiency of siRNA knockdown was determined by western blotting with PP6R3 and PP4R2 antibodies as indicated. Data are mean ± s.e.m. *P* values determined by one-way ANOVA. Bottom, representative images of the telomere FISH experiments. The arrowheads show loss of telomere signal. Red, telomere PNA FISH; blue, DAPI. **f**, Phi29-dependent telomere circles (top) detected in cells as indicated in **e**. The extent of [^{32}P] incorporation was quantified (bottom) from the autoradiographs, and the level of [^{32}P] incorporation by cells transfected with control siRNA was arbitrarily assigned a value of 100%. Data are mean ± s.d. and from two independent experiments. *P* values determined by one-way ANOVA. In **a**–**f**, the experiments were independently repeated at least twice with similar results.



Extended Data Fig. 6 | Replication defects in *Trf2*$^{F/-}$ MEFs in the absence of TRF2 phosphorylation at Ser365 or Ser367. a, Quantification of global replication fork dynamics (left) and rates of replication fork progression (right) of the IdU/CldU double pulse-labelling experiment in *Trf2*$^{F/-}$ MEFs complemented with empty vector, wild-type or mutant TRF2, performed 96 h after infection with control- or Cre-expressing adenovirus (*n* denotes number of analysed forks). Data are mean ± s.e.m. of triplicate experiments. Box plots are as in Fig. 3e. **b**, Representative images of the experiment from **a**. **c**–**e**, Quantification of micronuclei (**c**; 500 nuclei per replicate), mitotic catastrophe (**d**; 500 nuclei per replicate), and 53BP1 foci frequency (**e**; 150 nuclei per replicate) in *Trf2*$^{F/-}$ MEFs complemented as in **a**. Data are mean ± s.e.m. of three (**c**, **d**) or two (**e**) independent experiments. **f**, DNA damage in *Trf2*$^{F/-}$ MEFs complemented as in **a** was estimated by counting the frequency of cells with five or more 53BP1 foci. For each independent experiment ($n = 2$), a minimum of 150 nuclei of each condition were analysed. All *P* values were determined by one-way ANOVA.



Extended Data Fig. 7 | Suppression of constitutive binding of RTEL1 to the TRF2(S367A) phospho-dead mutant rescues replication defects in MEFs. **a**, Quantification of rates of replication fork progression (left) and representative images (right) of the IdU/CldU double pulse-labelling experiment in double-knockout *Trf2*$^{F/F}$;*Rtel1*$^{F/F}$ mouse ear fibroblasts stably expressing Myc–TRF2(S367A), together with wild-type V5–RTEL1 (WT) or C4C4 mutant V5–RTEL1(R1237H) (R/H), performed 96 h after infection with Cre-expressing adenovirus. Data are mean ± s.e.m. of triplicate experiments. **b**, Quantification of replication fork dynamics (top) and fork asymmetry (bottom) from cells as in **a**. Staining with anti-ssDNA antibody (right) was used to exclude broken DNA tracks (*n* denotes number of analysed forks). Box plots are as in Fig. 3e. Data are mean ± s.e.m. **c**–**e**, Quantification of the frequency of micronuclei (**c**; 500 nuclei per replicate), mitotic catastrophe (**d**; 500 nuclei per replicate), and 53BP1 foci (**e**; 150 nuclei per replicate) in *Trf2*$^{F/F}$;*Rtel1*$^{F/F}$ mouse ear fibroblasts complemented as indicated in **a**. Data are mean ± s.e.m. of three (**c**, **d**) or two (**e**) independent experiments. **f**, DNA damage in *Trf2*$^{F/F}$;*Rtel1*$^{F/F}$ mouse ear fibroblasts complemented as in **a** was estimated by counting the frequency of cells with five or more 53BP1 foci. For each independent experiment ($n = 2$), a minimum of 150 nuclei of each condition were analysed. All *P* values were determined by one-way ANOVA.



Extended Data Fig. 8 | TRF2(S367A) mutation induces TIFs and impairs formation of t-loops. a, Quantification (top) of the cross-linking efficiency test (bottom) in the *Trf2*$^{F/F}$ MEFs stably expressing wild-type or S367A mutant Myc–TRF2 120 h after treatment with 4-OHT ($n = 3$ independent biological replicates). Data are mean ± s.e.m. **b**, Left, quantification of TIFs per interphase in *Trf2*$^{F/-}$ MEFs complemented with empty vector, wild-type Myc–TRF2, phospho-dead mutant TRF2(S367A), or phospho-mimetic Myc–TRF2(S367D) and Myc–TRF2(S367E) mutants 96 h after infection with Cre-expressing adenovirus. Right, representative interphase TIF images. **c**, Quantification of TIFs per interphase in *Trf2*$^{F/F}$;*Rtel1*$^{F/F}$ mouse ear fibroblasts stably expressing Myc–TRF2(S367A) together with wild-type V5–RTEL1 or mutant V5–RTEL1(R1237H) 96 h after infection with GFP- or Cre-expressing adenovirus. **d**, Quantification (left) and representative images (right) of RPA staining at TIFs in *Trf2*$^{F/F}$;*Rtel1*$^{F/F}$ mouse ear fibroblasts complemented as in **c**. Analysis was carried out 96 h after infection with Cre-expressing adenovirus. Data in **b**–**d** are mean ± s.d. from three independent experiments ($n = 100$ cells in each treatment group analysed per independent experiment). **e**, Measurement of linear and t-loop molecules shown in Fig. 4c ($n = 3$ biological replicates scoring ≥1,192 molecules per replicate). T-loop measurements are a sum of the loop and tail portions of the molecule. Data are mean ± s.e.m. **f**, Measurement of the loop portion of t-loops from the experiments depicted in Fig. 4c ($n = 3$ biological replicates scoring ≥1,192 molecules per replicate). All *P* values were determined by one-way ANOVA.

nature research

Corresponding author(s): Simon J. Boulton

Last updated by author(s): Sep 13, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	N/A
Data analysis	N/A

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data generated or analyzed during this study are included in this article

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was chosen based on the experimental approach we adopted best practice according to the published literature.
Data exclusions	N/A
Replication	Every experiment was repeated at least 2 or 3 times.
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	This is all included in the Supplementary Information (Table 1)
Validation	The antibodies used were either validated based on previous publications or were confirmed by knock-down or mutation of the epitope.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	This is described in the materials and methods
Authentication	This is described in the materials and methods
Mycoplasma contamination	All cell lines used are routinely tested for mycoplasma as part of our institute policy.
Commonly misidentified lines (See ICLAC register)	N/A

Quantifying secondary transport at single-molecule resolution

https://doi.org/10.1038/s41586-019-1747-5

Received: 5 January 2019

Accepted: 7 October 2019

Published online: 13 November 2019

Gabriel A. Fitzgerald[1,6], Daniel S. Terry[1,2,6], Audrey L. Warren[3], Matthias Quick[3,4]*, Jonathan A. Javitch[3,4,5]* & Scott C. Blanchard[1,2]*

Secondary active transporters, which are vital for a multitude of physiological processes, use the energy of electrochemical ion gradients to power substrate transport across cell membranes[1,2]. Efforts to investigate their mechanisms of action have been hampered by their slow transport rates and the inherent limitations of ensemble methods. Here we quantify the activity of individual MhsT transporters, which are representative of the neurotransmitter:sodium symporter family of secondary transporters[3], by imaging the transport of individual substrate molecules across lipid bilayers at both single- and multi-turnover resolution. We show that MhsT is active only when physiologically oriented and that the rate-limiting step of the transport cycle varies with the nature of the transported substrate. These findings are consistent with an extracellular allosteric substrate-binding site that modulates the rate-limiting aspects of the transport mechanism[4,5], including the rate at which the transporter returns to an outward-facing state after the transported substrate is released.

Secondary active transporters are integral membrane proteins that use electrochemical ion gradients to translocate substrates across cellular membranes to subserve numerous physiological functions, which range from nutrient uptake, ion homeostasis and antimicrobial efflux to neurotransmission in humans[6,7]. Neurotransmitter:sodium symporters (NSS) modulate synaptic activity by clearing the synapse of neurotransmitters[7]. Despite their physiological and medical importance as molecular targets for therapeutic and recreational drugs[8,9], the individual steps of the transport cycle—and their regulation by neuromodulatory agents—remain poorly understood.

Ensemble electrophysiological, radiotracer and fluorometry studies have shed light on the transport mechanism for specific secondary active transporters, including those in the NSS family[10–15]. Such measurements necessarily infer the rates of specific steps in the transport cycle using complex and potentially disruptive biochemical or physiological manipulations. Single-molecule methods have previously been used to examine dynamics in individual transporters associated with key conformational states visited during the transport cycle[16–21]. However, no method yet exists to monitor directly solute transport by secondary transporters at the single-molecule level[22–25].

Here we report an approach based on single-molecule fluorescence resonance energy transfer (smFRET) that provides precise measurements of the activities of single transporters, and which has the capacity to detect individual steps in the NSS transport cycle. This approach does not involve modifications of the transporter itself but instead reports the movement of the transported substrate molecules into the lumen of proteoliposomes, using an encapsulated sensor protein[26–28]. We demonstrate the utility of this method by investigating the prokaryotic NSS homologue MhsT, a hydrophobic amino acid transporter that exhibits transport rates similar to those of human NSS proteins[3]. Single-molecule studies revealed that MhsT efficiently catalyses substrate transport only when the transporter is in the physiological orientation in the membrane. Single-turnover measurements revealed the kinetics of the first half-cycle of transport, which includes substrate binding, transporter isomerization and substrate release. Comparative kinetic analyses of single- and multi-turnover transport rates delineated the kinetics of the second half-cycle, which returns the transporter to the outward-facing state after release of substrate into the lumen. These measurements revealed that the rate of the return step, in which the transporter is classically modelled as devoid of substrate, is dependent on the identity of the substrate that is transported in the first half of the transport cycle. Consistent with these findings, we found that the time-averaged occupancy of the inward open state of MhsT in living cells was dependent on the identity of the substrate being transported. These insights provide compelling evidence that an allosteric ligand-binding site in the extracellular vestibule of MhsT[5]—and potentially also in related NSS homologues[4,29]—contributes to both the function and regulation of the transport cycle.

Engineered amino acid sensors

Bacterial periplasmic binding proteins help to support cellular growth and proliferation by scavenging substrates from the environment of the cell using a clamshell-like closure mechanism[26–28,30,31]. To enable the detection of ligand binding via FRET, we attached donor (LD550) and acceptor (LD650)[32,33] fluorophores in a site-specific manner within the opposing domains of the leucine, isoleucine, valine periplasmic binding protein (LIV-BP)[27] of *Escherichia coli* via the engineered cysteine

[1]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. [2]Department of Structural Biology, St. Jude Children's Research Hospital, Memphis, TN, USA. [3]Department of Psychiatry, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA. [4]Division of Molecular Therapeutics, New York State Psychiatric Institute, New York, NY, USA. [5]Department of Pharmacology, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA. [6]These authors contributed equally: Gabriel A. Fitzgerald, Daniel S. Terry. *e-mail: mq2102@cumc.columbia.edu; jaj2@cumc.columbia.edu; Scott.Blanchard@stjude.org

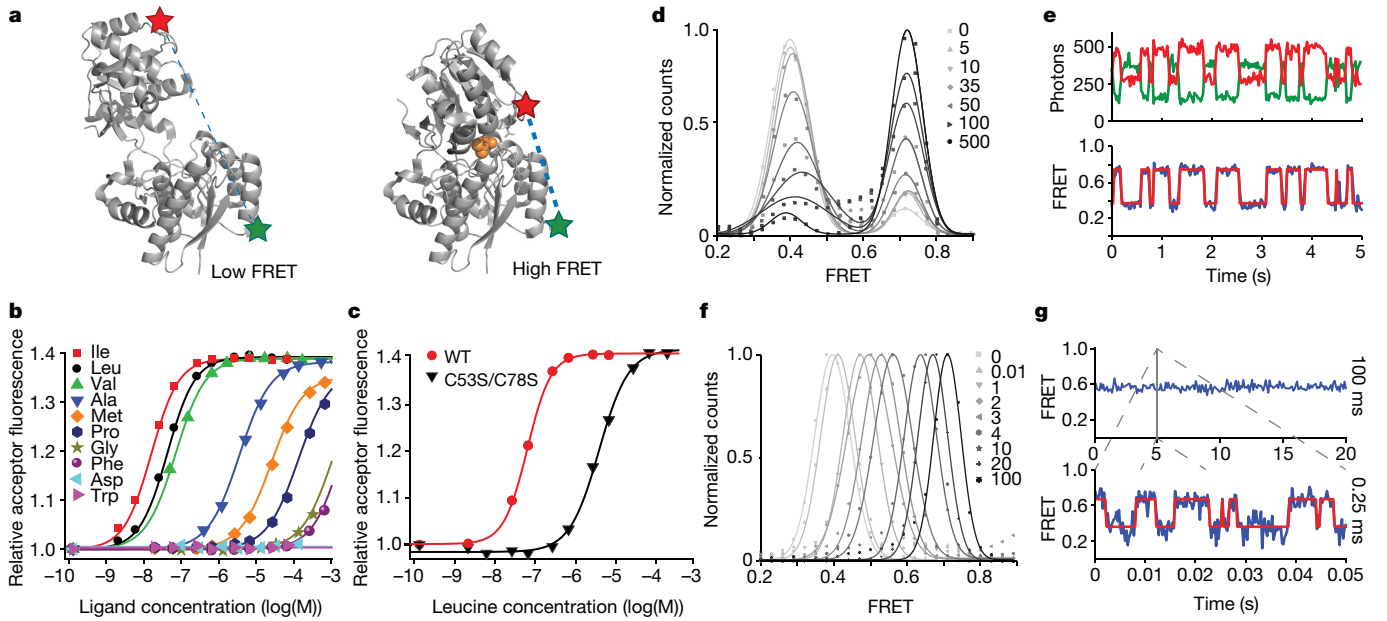


Fig. 1 | Design and characterization of a hydrophobic amino acid sensor. **a**, Apo (left) (RCSB Protein Data Bank code (PDB) 1Z15) and substrate (orange)-bound (right) (PDB 1Z16) crystal structures of LIV-BP. Locations of fluorophores are indicated by stars. **b**, Acceptor fluorescence of LIV-BP^{WT} under donor excitation was measured in the presence of the indicated concentrations of isoleucine, leucine, valine, alanine, methionine, phenylalanine, glycine, proline, asparagine and tryptophan. **c**, Acceptor fluorescence of LIV-BP^{WT} (WT) and LIV-BP^{SS} (C53S/C78S) in the presence of the indicated concentrations of leucine. Lines are fits to dose–response functions (Supplementary Table 1). **d**, FRET distributions at a range of nanomolar leucine concentrations for LIV-BP^{WT}. **e**, Single molecule trace of LIV-BP^{WT} at 40 nM leucine and 25-ms time resolution. **f**, FRET distributions at a range of micromolar leucine concentrations for LIV-BP^{SS}. **g**, Single molecule traces of LIV-BP^{SS} at 3.8 μM leucine and 100-ms (top) or 0.25-ms time resolution (bottom). All experiments were repeated at least three times with similar results.

residues N67C and E181C (hereafter, LIV-BP^{WT}) (Fig. 1a, Supplementary Methods).

We first characterized the binding properties of LIV-BP^{WT} using bulk fluorimetry methods (Supplementary Methods) and observed a robust 40% increase in acceptor fluorescence upon addition of saturating ligand concentrations (Fig. 1b), consistent with the expected distance change upon ligand binding[31]. LIV-BP^{WT} bound leucine, isoleucine and valine with nanomolar affinity, exhibiting dissociation constants (K_D) (approximately 40, 15, and 70 nM, respectively) that were independent of buffer conditions (Fig. 1b, Extended Data Fig. 1, Supplementary Table 1).

Given that 1 substrate molecule within the lumen of a 100-nm vesicle equates to approximately 3.5 μM[34–36], a single substrate molecule is expected to saturate a single encapsulated LIV-BP^{WT} sensor (Supplementary Note 1). To enable the detection of multiple transport cycles, we engineered a lower-affinity sensor in which the two native cysteines within LIV-BP were mutated to serine residues (C53S and C78S) (hereafter, LIV-BP^{SS}). The LIV-BP^{SS} sensor displayed similar ligand-dependent increases in FRET efficiency, but exhibited more than 100-fold decreases in leucine, isoleucine and valine binding affinities (5–30 μM) (Fig. 1c, Supplementary Table 1).

Analogue and digital amino acid sensing

We evaluated the rates of leucine binding to the LIV-BP^{WT} and LIV-BP^{SS} sensors by imaging the FRET behaviours of thousands of individual, surface-immobilized sensors simultaneously using wide-field, prism-based total-internal reflection fluorescence imaging[37,38] (Supplementary Methods). In the absence of ligands and at a 100-ms time resolution, both sensors displayed a mean FRET efficiency of about 0.4 and exhibited few (if any) spontaneous fluctuations in FRET (Extended Data Fig. 2a, b). In the presence of leucine, LIV-BP^{WT} displayed discrete transitions to a higher FRET state (mean FRET efficiency of about 0.70), the occupancy of which increased as a function of substrate concentration (Fig. 1d, e). By contrast, the LIV-BP^{SS} sensor displayed a single peak in the FRET histogram at all ligand concentrations, the mean value of which increased in an analogue fashion, reaching a plateau at about 0.70 FRET at concentrations more than 10-fold above its K_D (approximately 50 μM) (Fig. 1f). Hypothesizing that this analogue FRET response reflected the time-averaging of rapid FRET fluctuations between open and closed states, we increased the time resolution of the imaging to below 1 ms, at which we observed the LIV-BP^{SS} sensor to exhibit two clearly defined low- and high-FRET states with absolute values that were indistinguishable from those exhibited by LIV-BP^{WT} (Fig. 1g, Extended Data Fig. 2c, d). Detailed kinetic analyses revealed that the digital- and analogue-like FRET response behaviours of the LIV-BP^{WT} and LIV-BP^{SS}, respectively, reflect underlying distinctions in substrate off-rates at near-diffusion-limited substrate association rates (Extended Data Figs. 2e–j, 3a–c, Supplementary Note 2).

Single-molecule transport by MhsT

To measure leucine transport mediated by MhsT, we used a biotin–(NTA-Ni^{2+}) bridge[39] to surface-immobilize 100-nm proteoliposomes that were reconstituted under conditions that minimized the incorporation of more than 1 MhsT protein and LIV-BP^{SS} sensor in any single proteoliposome (Fig. 2a, Supplementary Methods). The encapsulation of LIV-BP^{SS} had no measurable effect on sensor activity (Extended Data Fig. 3i, j). The subpopulation of proteoliposomes that contained both a single MhsT and a single LIV-BP^{SS} was computationally selected on the basis of MhsT transport activity.

Pre-steady-state transport measurements were initiated in surface-immobilized proteoliposomes by rapidly exchanging the external buffer solution with different pH, Na^+ and substrate concentrations, while imaging at 100-ms time resolution (approximately 10 times faster than the anticipated turnover rate[3]). Accurate demarcation of transport

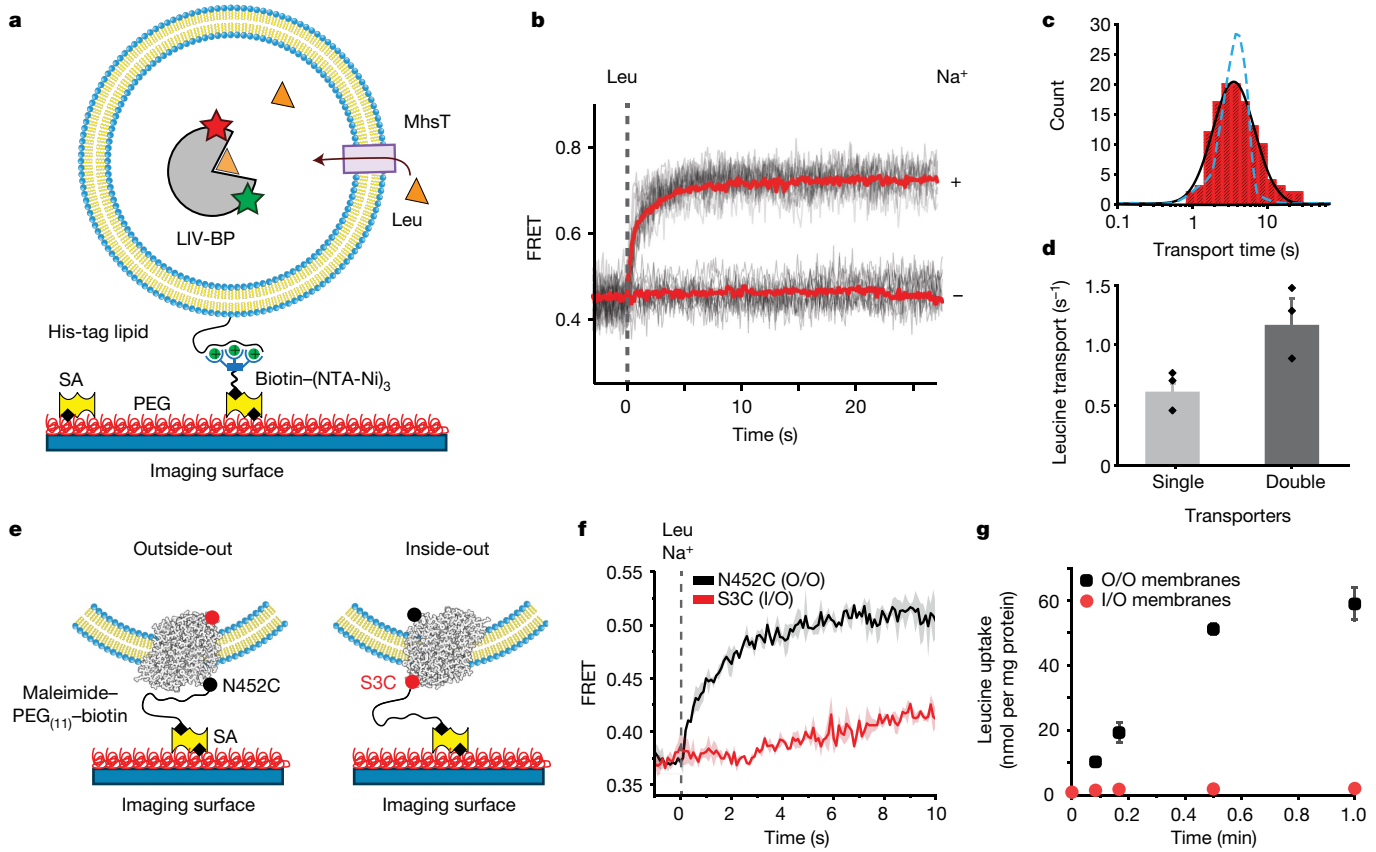


Fig. 2 | Single-molecule, multi-turnover MhsT transport assay. a, Schematic of the single-molecule transport assay with liposomes containing lipids functionalized with 6× histidine residues and immobilized onto a PEG-passivated glass surface via streptavidin (SA)–biotin–$(NTA\text{-}Ni^{2+})_3$ linker. **b**, Representative responding smFRET traces (grey) and ensemble-averaged FRET (red) from experiments imaging MhsT-containing proteoliposomes (pH 6 inside) with encapsulated LIV-BP^{SS} before and after (vertical dashed line) the external addition of 0.5 μM leucine and either 150 mM (+) or 0 mM (–) Na^+ at pH 8. This experiment was repeated at least three times with similar results. **c**, Histogram of single exponential time constants from individual FRET traces (bars) fit to a log-normal distribution (black line). Simulated traces approximate the variation in rates caused by the liposome size distribution (blue dashed line). **d**, Transport rate of liposomes containing one (light grey) or two (dark grey) labelled MhsT transporters. Similar results were observed over three experiments. **e**, Schematic of MhsT proteoliposomes immobilized in the outside-out (left) or inside-out (right) orientations. **f**, Average FRET curves of transporters immobilized in the inside-out (I/O) orientation (S3C, red) and outside-out (O/O) orientation (N452C, black) before and after (vertical dashed line) the external addition of 2.5 μM leucine and 150 mM Na^+ at pH 8. Mean ± s.e.m. represented by shaded area ($n = 4$ experiments). **g**, Radioactive leucine uptake using membrane vesicles of *E. coli* cells expressing MhsT in outside-out (black) and inside-out (red) orientations. Data points represent mean ± s.e.m. ($n = 3+$ experiments).

initiation was established by co-injecting a fluorescently labelled tracer (Extended Data Fig. 4, Supplementary Methods). Using an external solution containing 0.5 μM leucine and 150 mM Na^+, we observed that individual LIV-BP^{SS} sensors underwent a rapid increase in FRET that reached a plateau at a value that was consistent with leucine saturation (Fig. 2b, Extended Data Fig. 5a, c–e). No such increase was observed in the absence of Na^+ (Fig. 2b). These data are consistent with the observed FRET change that arises from an encapsulated FRET sensor responding to a Na^+-dependent concentrative translocation of leucine into the proteoliposome lumen over multiple transport cycles.

We characterized the distribution of transport rates within the population of liposomes by fitting the individual FRET traces of FRET-responsive proteoliposomes to single exponential functions (Extended Data Fig. 5a). The distribution of time constants observed was well-described by a log-normal distribution (Fig. 2c); this is consistent with a relatively uniform transport rate of monomeric MhsT molecules, the variance of which closely recapitulated the expected distribution of extruded proteoliposomes sizes[34–36] (Fig. 2c). We therefore compiled all FRET-responsive proteoliposomes within each experiment to obtain a mean transport rate for the ensemble (red line in Fig. 2b), which we then converted to luminal substrate concentrations (Extended Data Fig. 1b, c).

We confirmed that over 90% of surface-immobilized, responding proteoliposomes contained only a single MhsT transporter using a fully active Cy7-labelled MhsT construct (Extended Data Figs. 6a, 7). The rate of Cy7-labelled MhsT transport closely mirrored that of the unlabelled molecule (Fig. 2d, Extended Data Fig. 9b). As expected, the minor subpopulation of proteoliposomes that contained two Cy7-labelled MhsT transporters exhibited transport rates that were approximately twofold faster than those with a single transporter (Fig. 2d).

MhsT transport is orientation-dependent

The reconstitution of purified integral membrane proteins into proteoliposomes can give rise to both inside-in and inside-out orientations of the protein[40–42]. The unimodal distribution of transport rates observed therefore suggests that: (1) only one MhsT orientation is present; (2) only one orientation is active; or (3) both MhsT orientations are present and transport at similar rates. To distinguish between these models, we immobilized 100-nm proteoliposomes reconstituted with MhsT mutants labelled with biotin–11-unit polyethylene glycol ($PEG_{(11)}$)–maleimide, which retained full transport activity (Extended Data Fig. 6a, b), to isolate inside-out (S3C mutant, residue facing the cytoplasm) and outside-out (N452C mutant, residue facing extracellularly) orientations

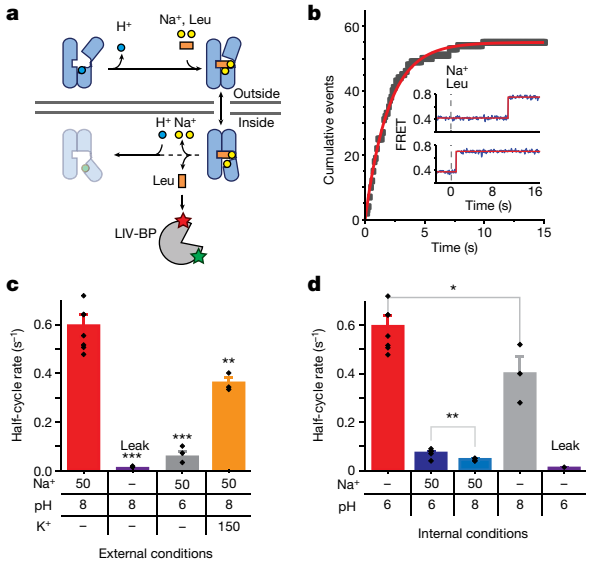


Fig. 3 | Single-molecule, single-turnover measurements of first half-cycle MhsT transport rates. a, Schematic of the single-turnover assay. **b**, Cumulative distribution (grey squares) from one replicate fit with a single exponential (red line) in 50 mM external Na^+ at pH 8 and 0.1 μM leucine and internal pH 6. Similar results were observed over many experiments. Inset, representative smFRET traces (blue) and idealization (red) from experiments imaging MhsT proteoliposomes with encapsulated LIV-BP^{WT} before and after the addition of 0.1 μM leucine and 50 mM Na^+ (grey dashed line). Traces are representative of experiments performed at least three times. **c**, First half-cycle rates with varying external ion and pH conditions with 0.1 μM leucine and pH 6 inside, demonstrating that all three conditions are statistically significantly different from the basal conditions (red). $^{**}P = 0.014$, $^{***}P < 0.0001$. Leak, rate of passive substrate leak across the lipid bilayer. **d**, First half-cycle rates with varying internal conditions with constant external conditions: 0.1 μM leucine and 50 mM Na^+ at pH 8. $^{*}P = 0.048$, $^{**}P = 0.019$. The red and purple bars in **c** and **d** are duplicated for comparison. Black diamonds represent individual data points, and bars represent mean ± s.e.m. ($n = 3+$ experiments). Significance was tested with a two-sided Student's *t*-test that was not corrected for multiple hypotheses.

(Fig. 2e). As expected from proteoliposome immobilization via the embedded transporter, we observed a 5–10-fold enrichment in the number of proteoliposomes that were transport-active (Extended Data Fig. 8a).

Proteoliposomes bearing a single outside-out oriented MhsT protein exhibited transport rates that matched those found when proteoliposomes were immobilized via a Ni^{2+}-NTA bridge (Fig. 2f, Extended Data Fig. 8a, Supplementary Methods). By contrast, proteoliposomes bearing a single inside-out oriented MhsT exhibited no detectable transport activity (Fig. 2f, Extended Data Fig. 8a). We found similar results when proteoliposomes that contain MhsT were immobilized using two additional cytoplasmic-facing residues, P86C and G87C (Extended Data Fig. 8b, c); these mutants also retained full activity in bulk mixed-orientation assays (Extended Data Fig. 6b). We verified that inside-out oriented MhsT transporters are effectively non-functional in transport by preparing membrane vesicles of defined orientation from *E. coli* heterologously expressing MhsT[41,43] (Fig. 2g, Extended Data Fig. 6c, Supplementary Methods). By contrast, an unrelated secondary active glycine transporter that is found natively in bacterial membranes[44,45] was functional in membrane vesicles of both orientations (Extended Data Fig. 6d). We therefore conclude that MhsT efficiently catalyses transport only in the physiological (outside-out) orientation, and that proteoliposomes that bear inside-out oriented MhsT do not contribute to our transport measurements.

Multi-turnover MhsT transport kinetics

We next set out to determine the kinetic parameters of leucine transport by MhsT using the smFRET transport assay. In good agreement with ensemble measurements of the uptake of radiolabelled ligand, we found that the mean initial rates of MhsT transport varied as a function of external leucine concentration (Extended Data Fig. 9a, b) and that the catalytic turnover rate (k_{cat}) was approximately 1.18 ± 0.04 s^{-1} and the Michaelis–Menten constant (K_m) was approximately 0.06 ± 0.01 μM (Extended Data Fig. 9b). As expected, MhsT also exhibited transport rates that were strongly dependent on external Na^+ concentration with an apparent K_m of 53 ± 4 mM (Extended Data Fig. 9c). Consistent with MhsT being a proton antiporter[46], MhsT transport rates increased at higher external pH, and decreased with higher internal pH (Extended Data Figs. 9d, 10).

Single-turnover MhsT transport

To investigate the underlying mechanistic features of the secondary transport process, we isolated the first half-cycle of transport (from external substrate binding to internal substrate release) by performing analogous experiments with the LIV-BP^{WT} sensor encapsulated in the lumen of a 100-nm proteoliposome, which is saturated after only a single transport event (Fig. 3a, Supplementary Note 1). Consistent with this expectation, the rapid addition of leucine and Na^+ to surface-immobilized proteoliposomes that contain a single encapsulated LIV-BP^{WT} triggered step-like increases in FRET from baseline (FRET of about 0.40) to that of a fully saturated sensor (FRET of about 0.70) (Fig. 3b inset, Extended Data Fig. 5b). Notably, such events occurred at variable time delays after the initiation of transport. As substrate mixing and substrate binding to the sensor are both rapid processes (about 100 ms) (Extended Data Figs. 3e–h, 4), we reasoned that the observed delay reflects the total time required for the transporter to bind the substrate in an outward-facing conformation, isomerize to the inward-facing state and release substrate into the proteoliposome lumen (Fig. 3a, b). After accounting for passive substrate leak across the lipid bilayer, which was two orders of magnitude slower than transport (Figs. 3c, 4a), we fit the cumulative distribution of the delay times between transport initiation and the observed FRET event to a single exponential process to estimate the time required for MhsT to transit its first half-cycle (Fig. 3b).

At subsaturating external substrate concentrations (0.1 μM leucine and 50 mM Na^+) and an outwardly directed pH gradient (pH 6 inside and pH 8 outside), we estimated the first half-cycle transport rate of MhsT to be about 0.62 ± 0.08 s^{-1} (Fig. 3c, Supplementary Note 1). As expected, substrate accumulation was not observed in the absence of Na^+ (Fig. 3c). The rate of single-turnover transport decreased by nearly an order of magnitude when the external pH was increased from 6 to 8 (Fig. 3c). We infer from these findings that high external proton concentrations reduce Na^+ and substrate binding to MhsT.

In the related NSS-family transporter LeuT, internal Na^+ release is thought to precede substrate dissociation[4,21]. We therefore reasoned that high luminal concentrations of Na^+ should reduce the first half-cycle transport rate. Consistent with this model, we found that luminal Na^+ (50 mM) decreased substrate release from MhsT by nearly an order of magnitude (Fig. 3d). When present at equal concentrations (50 mM) both inside and outside, residual transport remained approximately tenfold faster than passive leucine leak across the bilayer (Fig. 3d). Lowering the luminal proton concentration decreased the apparent first half-cycle transport rate even further (Fig. 3d). These data are consistent with Na^+-facilitated substrate diffusion through MhsT (Extended Data Fig. 9e) and indicate that luminal protons can influence the interaction between Na^+ and the transporter to modulate substrate release from the intracellular surface of MhsT.

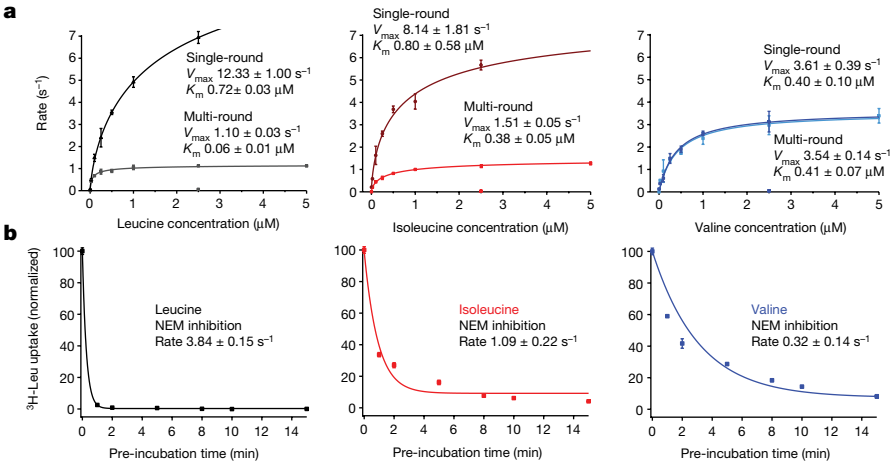


Fig. 4 | Substrate identity affects the return rate of MhsT after substrate release to the proteoliposome lumen. a, Substrate titrations in 150 mM Na^+ at pH 8 of both single-round (darker line) and multi-round (lighter line) transport assays for leucine (left), isoleucine (middle) and valine (right), each fit with a Michaelis–Menten equation to obtain kinetic parameters as in Extended Data Fig. 9b for multi round and Fig. 3b for the single round. Individual points at 2.5 µM indicate the rate of translocation in the absence of transporter in the single-round assay. Leak rates were 0.027 s^{-1}, 0.025 s^{-1} and 0.015 s^{-1} for leucine, isoleucine and valine, respectively. **b**, Normalized radioactive substrate uptake over 3 s by *E. coli* heterologously expressing MhsT (A22C) after pre-incubation for the indicated time with either leucine (left), isoleucine (middle) or valine (right) and NEM and fit with single exponentials. Data points represent mean ± s.e.m. (n = 3+ experiments).

To examine whether Na^+ leaves before H^+ binding or whether H^+ binds to the fully loaded transporter to actively facilitate Na^+ release, we increased the luminal pH in the absence of luminal Na^+. Here the presence of lower luminal proton concentration significantly decreased the single-turnover transport rate (Fig. 3d). We conclude from these findings that protons accelerate the release of substrate from MhsT, presumably by facilitating Na^+ release via the protonation of residues in proximity to the Na^+-binding sites[46,47].

Rate-limiting steps in MhsT transport

Using the data from single- and multi-turnover transport assays, we set out to infer the rate of the second half of the transport cycle—the 'return step' that restores the inward-facing, substrate-unloaded transporter to an outward-facing state—by comparing the maximal rates of each assay. Using the LIV-BP^{WT} sensor, we measured the maximal first half-cycle leucine-transport rate (k_{cat}) to be about 12.0 ± 0.10 s^{-1} (Fig. 4a). By contrast, using the LIV-BP^{SS} sensor, the k_{cat} of multi-turnover leucine-transport rate was about 1.10 ± 0.03 s^{-1} (Fig. 4a). This notable distinction in single-turnover and multi-turnover rates implies that the return step, estimated here to occur at a rate of approximately 1 s^{-1}, is rate-limiting to the overall transport cycle. Previous investigations have suggested that the return step after substrate release is rate-limiting to the NSS transport mechanism[48–50]. However, this suggestion appears to be incompatible with data that show that transporters (such as the human serotonin transporter) catalyse the transport of distinct substrates (such as dopamine and serotonin) at markedly different rates[51].

To determine how different substrates affect rate-limiting steps of transport in MhsT, we took advantage of the broad specificity of MhsT for hydrophobic amino acids and the similar affinities of the LIV-BP sensors for leucine, isoleucine and valine (Fig. 1b, Extended Data Fig. 1c, Supplementary Table 1) to assess the single- and multi-turnover rates of these three substrates. In these experiments, MhsT exhibited first half-cycle and multi-turnover isoleucine-transport rates that were similar, but not identical, to those observed for leucine (k_{cat} of 8.14 ± 1.81 s^{-1} and 1.51 ± 0.05 s^{-1}, respectively) (Fig. 4a). These findings indicate that although the return step during isoleucine transport is somewhat faster than during leucine transport, this step is still rate-limiting for the full transport cycle. A nearly identical k_{cat} value for valine transport was observed with both single- and multi-turnover assays: 3.61 ± 0.39 and 3.54 ± 0.14 s^{-1}, respectively (Fig. 4a). This result implies that the return step is much faster, and no longer rate-limiting, when valine is the transported substrate. These findings could not be explained by differences in the rebinding rates of leucine, isoleucine and valine to the intracellular face of the transporter after luminal release (Extended Data Fig. 9f, Supplementary Note 4). These insights challenge canonical models of NSS transport in which the rate-limiting return step is defined as being substrate-free, and instead suggest that the rate-limiting return step in the MhsT transport cycle is modulated by substrate after the transported substrate has been released.

MhsT conformation in living cells

The variable first and second half-cycle reaction rates evidenced for different substrates implies that the time-averaged distribution of MhsT conformations should vary predictably with the identity of the transported substrate. To test this hypothesis, we performed live-cell sulfhydryl accessibility experiments in which the membrane-permeant alkylating agent *N*-ethylmaleimide (NEM) was used to probe the cytoplasmic accessibility of a cysteine residue (A22C) located within the translocation pathway of MhsT and the homologous NSS transporter, Tyt1[52]. In these experiments, more time spent, on average, in the inward-facing state should result in faster NEM reactivity. As NEM labelling at this position blocks transport, increased occupancy of the inward-facing state will result in more-rapid inhibition of transport activity. Consistent with the predictions of our single-molecule transport assays, we found that the rates of MhsT inhibition by NEM pretreatment were fastest when transporting leucine, followed by isoleucine and an order of magnitude slower when transporting valine (Fig. 4b). These findings support our conclusion that the rate of the return step in the transport mechanism depends on the identity of the substrate being transported.

Discussion

Here, using an unmodified transporter and substrates, we have established a fluorescence-based transport assay with single transporter and single substrate sensing resolution to reveal previously inaccessible aspects of the NSS transport mechanism. Using this approach, we observed that MhsT functions as a monomer, with a distribution of transport rates that is unimodal and homogenous. By leveraging the

capacity to isolate the first half-cycle of transport, we further showed that pH has a critical role in modulating substrate binding to, and release from, MhsT.

Determination of the full cycle and first half-cycle transport rates revealed that the return of the putatively unloaded transporter to the outward-facing state was dependent on the identity of the substrate being transported. This step has historically been difficult to discern experimentally, and to our knowledge has been inferred only from complex experiments and models[13–15,48–50,53–55]. It is conceivable that different substrates are released from different configurations of the transporter, with different rates of intracellular gate closing and return to the outward-facing state. Such a mechanism would require conformational memory exerting effects on the time scale of a second, the turnover rate that we observed. We find our results more consistent with reports of a second substrate (S2) binding site in the extracellular vestibule of MhsT[5] and the prokaryotic NSS homologue LeuT[4,56]. In this framework, substrate binds in the S2 site to allosterically trigger inward opening and Na^+ and S1 substrate release, but remains trapped until the transporter isomerizes to the outward open state, thereby modulating the rate of the return step. We verified our model in living cells through cysteine accessibility measurements, which demonstrate that the time-averaged occupancy of the inward open state of MhsT during active transport is substrate-dependent and directly correlated with the return-step rate. In principle, such conformational information could be used by the cell to respond to different nutrients in a manner akin to signalling proteins[57–61].

Notably, substrate transport from the extracellular face of MhsT to the intracellular face was robust, whereas the ability of MhsT to move substrate from the inside face to the outside face (whether in proteoliposomes or membrane vesicles) was negligible—despite favourable electrochemical gradients. Hence, in contrast to other secondary transporters, we conclude that MhsT can catalyse rapid substrate influx in the presence of appropriate electrochemical gradients, but not efflux, and that the S2 binding site at the extracellular face of the transporter is critical to vectorial substrate transport across the lipid bilayer.

Notably, amphetamine-induced dopamine efflux by the dopamine transporter, a mammalian NSS homologue, is dependent on phosphorylation of the N terminus of the transporter[42,62–64], which suggests that the transmembrane domain alone is inefficient at mediating efflux. Although the underlying mechanism is at present unknown, this extended N-terminal segment is absent in MhsT, which appears incapable of efflux.

The successful investigation of the secondary transport cycle using single-molecule techniques offers the potential to apply analogous methods to investigations of the multitude of primary and secondary transporters. Such methods may be particularly advantageous for biophysical investigations of mammalian transporters, which are more prone to biochemical and functional heterogeneities that compromise quantitative ensemble investigations. Applications of this kind are, in principle, limited only by the ability to functionally reconstitute transporters into proteoliposomes and the availability of sensors specific to the substrates being transported. The latter of these challenges can be overcome by enlisting the diverse array of periplasmic binding proteins for unique substrates that exists in nature[7,30], RNA aptamers[65], receptor soluble domains[66] or synthetic engineering or evolution efforts that can, in theory, be implemented to recognize any arbitrary small molecule[67,68].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

Code availability

The full source code of SPARTAN[37], which was used for all analysis of smFRET data, is publicly available at http://www.scottblanchardlab.com/software.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1747-5.

1. Forrest, L. R., Krämer, R. & Ziegler, C. The structural basis of secondary active transport mechanisms. *Biochim. Biophys. Acta* **1807**, 167–188 (2011).
2. Shi, Y. Common folds and transport mechanisms of secondary active transporters. *Annu. Rev. Biophys.* **42**, 51–72 (2013).
3. Malinauskaite, L. et al. A mechanism for intracellular release of Na^+ by neurotransmitter/sodium symporters. *Nat. Struct. Mol. Biol.* **21**, 1006–1012 (2014).
4. Shi, L., Quick, M., Zhao, Y., Weinstein, H. & Javitch, J. A. The mechanism of a neurotransmitter:sodium symporter-inward release of Na^+ and substrate is triggered by substrate in a second binding site. *Mol. Cell* **30**, 667–677 (2008).
5. Quick, M. et al. The LeuT-fold neurotransmitter:sodium symporter MhsT has two substrate sites. *Proc. Natl Acad. Sci. USA* **115**, E7924–E7931 (2018).
6. Pao, S. S., Paulsen, I. T. & Saier, M. H. Jr. Major facilitator superfamily. *Microbiol. Mol. Biol. Rev.* **62**, 1–34 (1998).
7. Kristensen, A. S. et al. SLC6 neurotransmitter transporters: structure, function, and regulation. *Pharmacol. Rev.* **63**, 585–640 (2011).
8. Sitte, H. H. & Freissmuth, M. Amphetamines, new psychoactive drugs and the monoamine transporter cycle. *Trends Pharmacol. Sci.* **36**, 41–50 (2015).
9. Sitte, H. H. et al. Carrier-mediated release, transport rates, and charge transfer induced by amphetamine, tyramine, and dopamine in mammalian cells transfected with the human dopamine transporter. *J. Neurochem.* **71**, 1289–1297 (1998).
10. Accardi, A. & Miller, C. Secondary active transport mediated by a prokaryotic homologue of ClC Cl^-channels. *Nature* **427**, 803–807 (2004).
11. Díez-Sampedro, A., Eskandari, S., Wright, E. M. & Hirayama, B. A. Na^+-to-sugar stoichiometry of SGLT3. *Am. J. Physiol. Renal Physiol.* **280**, F278–F282 (2001).
12. Gorraitz, E., Hirayama, B. A., Paz, A., Wright, E. M. & Loo, D. D. F. Active site voltage clamp fluorometry of the sodium glucose cotransporter hSGLT1. *Proc. Natl Acad. Sci. USA* **114**, E9980–E9988 (2017).
13. Mackenzie, B., Illing, A. C. & Hediger, M. A. Transport model of the human Na^+-coupled L-ascorbic acid (vitamin C) transporter SVCT1. *Am. J. Physiol. Cell Physiol.* **294**, C451–C459 (2008).
14. Eskandari, S. et al. Thyroid Na^+/I^- symporter. Mechanism, stoichiometry, and specificity. *J. Biochem.* **272**, 27230–27238 (1997).
15. Parent, L., Supplisson, S., Loo, D. D. & Wright, E. M. Electrogenic properties of the cloned Na^+/glucose cotransporter: II. A transport model under nonrapid equilibrium conditions. *J. Membr. Biol.* **125**, 63–79 (1992).
16. Majumdar, D. S. et al. Single-molecule FRET reveals sugar-induced conformational dynamics in LacY. *Proc. Natl Acad. Sci. USA* **104**, 12640–12645 (2007).
17. Zhao, Y. et al. Single-molecule dynamics of gating in a neurotransmitter transporter homologue. *Nature* **465**, 188–193 (2010).
18. Akyuz, N. et al. Transport domain unlocking sets the uptake rate of an aspartate transporter. *Nature* **518**, 68–73 (2015).
19. Akyuz, N., Altman, R. B., Blanchard, S. C. & Boudker, O. Transport dynamics in a glutamate transporter homologue. *Nature* **502**, 114–118 (2013).
20. Zhao, Y. et al. Substrate-modulated gating dynamics in a Na^+-coupled neurotransmitter transporter homologue. *Nature* **474**, 109–113 (2011).
21. Terry, D. S. et al. A partially-open inward-facing intermediate conformation of LeuT is associated with Na^+ release and substrate transport. *Nat. Commun.* **9**, 230 (2018).
22. Veshaguri, S. et al. Direct observation of proton pumping by a eukaryotic P-type ATPase. *Science* **351**, 1469–1473 (2016).
23. Watanabe, R. et al. Arrayed lipid bilayer chambers allow single-molecule analysis of membrane transporter activity. *Nat. Commun.* **5**, 4519 (2014).
24. Soga, N., Watanabe, R. & Noji, H. Attolitre-sized lipid bilayer chamber array for rapid detection of single transporters. *Sci. Rep.* **5**, 11025 (2015).
25. Watanabe, R., Soga, N., Ohdate, S.-Y. & Noji, H. Single-molecule analysis of membrane transporter activity by means of a microsystem. *Methods Mol. Biol.* **1700**, 321–330 (2018).
26. Okumoto, S. Quantitative imaging using genetically encoded sensors for small molecules in plants. *Plant J.* **70**, 108–117 (2012).
27. Trakhanov, S. et al. Ligand-free and -bound structures of the binding protein (LivJ) of the *Escherichia coli* ABC leucine/isoleucine/valine transport system: trajectory and dynamics of the interdomain rotation and ligand specificity. *Biochemistry* **44**, 6597–6608 (2005).
28. Seo, M.-H., Park, J., Kim, E., Hohng, S. & Kim, H.-S. Protein conformational dynamics dictate the binding affinity for a ligand. *Nat. Commun.* **5**, 3724 (2014).
29. Coleman, J. A. & Gouaux, E. Structural basis for recognition of diverse antidepressants by the human serotonin transporter. *Nat. Struct. Mol. Biol.* **25**, 170–175 (2018).
30. Scheepers, G. H., Lycklama a Nijeholt, J. A. & Poolman, B. An updated structural classification of substrate-binding proteins. *FEBS Lett.* **590**, 4393–4401 (2016).

31. Chino, S., Sakaguchi, A., Yamoto, R., Ferri, S. & Sode, K. Branched-chain amino acid biosensing using fluorescent modified engineered leucine/isoleucine/valine binding protein. *Int. J. Mol. Sci*. **8**, 513–525 (2007).
32. Altman, R. B. et al. Enhanced photostability of cyanine fluorophores across the visible spectrum. *Nat. Methods* **9**, 428–429 (2012).
33. Altman, R. B. et al. Cyanine fluorophore derivatives with enhanced photostability. *Nat. Methods* **9**, 68–71 (2012).
34. Korgel, B. A., van Zanten, J. H. & Monbouquette, H. G. Vesicle size distributions measured by flow field-flow fractionation coupled with multiangle light scattering. *Biophys. J*. **74**, 3264–3272 (1998).
35. Hope, M. J., Bally, M. B., Webb, G. & Cullis, P. R. Production of large unilamellar vesicles by a rapid extrusion procedure: characterization of size distribution, trapped volume and ability to maintain a membrane potential. *Biochim. Biophys. Acta* **812**, 55–65 (1985).
36. Hupfeld, S., Holsaeter, A. M., Skar, M., Frantzen, C. B. & Brandl, M. Liposome size analysis by dynamic/static light scattering upon size exclusion-/field flow-fractionation. *J. Nanosci. Nanotechnol*. **6**, 3025–3031 (2006).
37. Juette, M. F. et al. Single-molecule imaging of non-equilibrium molecular ensembles on the millisecond timescale. *Nat. Methods* **13**, 341–344 (2016).
38. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat. Methods* **5**, 507–516 (2008).
39. Lata, S., Reichel, A., Brock, R., Tampé, R. & Piehler, J. High-affinity adaptors for switchable recognition of histidine-tagged proteins. *J. Am. Chem. Soc*. **127**, 10205–10215 (2005).
40. Mulligan, C., Fitzgerald, G. A., Wang, D.-N. & Mindell, J. A. Functional characterization of a Na^+-dependent dicarboxylate transporter from *Vibrio cholerae*. *J. Gen. Physiol*. **143**, 745–759 (2014).
41. Quick, M., Tomasevic, J. & Wright, E. M. Functional asymmetry of the human Na^+/glucose transporter (hSGLT1) in bacterial membrane vesicles. *Biochemistry* **42**, 9147–9152 (2003).
42. Khoshbouei, H. et al. N-terminal phosphorylation of the dopamine transporter is required for amphetamine-induced efflux. *PLoS Biol*. **2**, e78 (2004).
43. Kaback, H. R. in *Enzyme Purification and Related Techniques (Methods in Enzymology vol. 22)* (ed. Jakoby, W. B.) 99–120 (Elsevier, 1971).
44. Wargel, R. J., Shadur, C. A. & Neuhaus, F. C. Mechanism of d-cycloserine action: transport systems for D-alanine, D-cycloserine, L-alanine, and glycine. *J. Bacteriol*. **103**, 778–788 (1970).
45. Akahane, S., Kamata, H., Yagisawa, H. & Hirata, H. A novel neutral amino acid transporter from the hyperthermophilic archaeon *Thermococcus* sp. KS-1. *J. Biochem*. **133**, 173–180 (2003).
46. Zhao, Y. et al. Substrate-dependent proton antiport in neurotransmitter:sodium symporters. *Nat. Chem. Biol*. **6**, 109–116 (2010).
47. Khelashvili, G. et al. Conformational dynamics on the extracellular side of LeuT controlled by Na^+ and K^+ ions and the protonation state of Glu^{290}. *J. Biol. Chem*. **291**, 19786–19799 (2016).
48. Schicker, K. et al. Unifying concept of serotonin transporter-associated currents. *J. Biol. Chem*. **287**, 438–445 (2012).
49. Jones, S. R., Joseph, J. D., Barak, L. S., Caron, M. G. & Wightman, R. M. Dopamine neuronal transport kinetics and effects of amphetamine. *J. Neurochem*. **73**, 2406–2414 (1999).
50. Erreger, K., Grewer, C., Javitch, J. A. & Galli, A. Currents in response to rapid concentration jumps of amphetamine uncover novel aspects of human dopamine transporter function. *J. Neurosci*. **28**, 976–989 (2008).
51. Larsen, M. B. et al. Dopamine transport by the serotonin transporter: a mechanistically distinct mode of substrate translocation. *J. Neurosci*. **31**, 6605–6615 (2011).
52. Quick, M. et al. State-dependent conformations of the translocation pathway in the tyrosine transporter Tyt1, a novel neurotransmitter:sodium symporter from *Fusobacterium nucleatum*. *J. Biol. Chem*. **281**, 26444–26454 (2006).
53. Zomot, E. et al. Mechanism of chloride interaction with neurotransmitter:sodium symporters. *Nature* **449**, 726–730 (2007).
54. Mager, S. et al. Ion binding and permeation at the GABA transporter GAT1. *J. Neurosci*. **16**, 5405–5414 (1996).
55. Bicho, A. & Grewer, C. Rapid substrate-induced charge movements of the GABA transporter GAT1. *Biophys. J*. **89**, 211–231 (2005).
56. Quick, M. et al. Binding of an octylglucoside detergent molecule in the second substrate (S2) site of LeuT establishes an inhibitor-bound conformation. *Proc. Natl Acad. Sci. USA* **106**, 5563–5568 (2009).
57. Harwood, C. S., Nichols, N. N., Kim, M. K., Ditty, J. L. & Parales, R. E. Identification of the pcaRKF gene cluster from *Pseudomonas putida*: involvement in chemotaxis, biodegradation, and transport of 4-hydroxybenzoate. *J. Bacteriol*. **176**, 6479–6488 (1994).
58. Thevelein, J. M. et al. Novel mechanisms in nutrient activation of the yeast protein kinase A pathway. *Acta Microbiol. Immunol. Hung*. **55**, 75–89 (2008).
59. Thevelein, J. M. & Voordeckers, K. Functioning and evolutionary significance of nutrient transceptors. *Mol. Biol. Evol*. **26**, 2407–2414 (2009).
60. Goberdhan, D. C. I., Wilson, C. & Harris, A. L. Amino acid sensing by mTORC1: intracellular transporters mark the spot. *Cell Metab*. **23**, 580–589 (2016).
61. Scalise, M. et al. Insights into the transport side of the human SLC38A9 transceptor. *Biochim. Biophys. Acta* **1861**, 1558–1567 (2019).
62. Pizzo, A. B. et al. The membrane raft protein Flotillin-1 is essential in dopamine neurons for amphetamine-induced behavior in *Drosophila*. *Mol. Psychiatry* **18**, 824–833 (2013).
63. Karam, C. S. & Javitch, J. A. Phosphorylation of the amino terminus of the dopamine transporter: regulatory mechanisms and implications for amphetamine action. *Adv. Pharmacol*. **82**, 205–234 (2018).
64. Pizzo, A. B. et al. Amphetamine-induced behavior requires CaMKII-dependent dopamine transporter phosphorylation. *Mol. Psychiatry* **19**, 279–281 (2014).
65. Wang, R. E., Zhang, Y., Cai, J., Cai, W. & Gao, T. Aptamer-based fluorescent biosensors. *Curr. Med. Chem*. **18**, 4175–4184 (2011).
66. Marvin, J. S. et al. An optimized fluorescent probe for visualizing glutamate neurotransmission. *Nat. Methods* **10**, 162–170 (2013).
67. Marcos, E. et al. Principles for designing proteins with cavities formed by curved β sheets. *Science* **355**, 201–206 (2017).
68. Tinberg, C. E. et al. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501**, 212–216 (2013).

Acknowledgements We thank R. B. Altman for the preparation of microscope slides and other reagents for single-molecule experiments, and M. Saper for providing the pET29a His-tagged LIV-BP plasmid. This work was supported by National Institutes of Health grants 1R21MH099491 and R01GM098859 (to S.C.B), DA041510 and U54 GM087519 (to J.A.J) and the National Institute on Drug Abuse of the National Institutes of Health grant F31DA044688-01 (to G.A.F).

Author contributions S.C.B. and D.S.T. conceptualized the single-molecule transport assay. G.A.F. prepared and labelled LIV-BP and carried out the single-molecule experiments. G.A.F., D.S.T. and S.C.B. analysed and interpreted the single-molecule data. J.A.J. and M.Q. designed the cysteine accessibility and vesicle topology studies and interpreted the data. M.Q. and A.L.W. prepared purified MhsT and MhsT-containing bacterial membranes. M.Q. carried out radioactive uptake assays and cysteine accessibility assays. All authors contributed to overall experimental design and the writing of the manuscript.

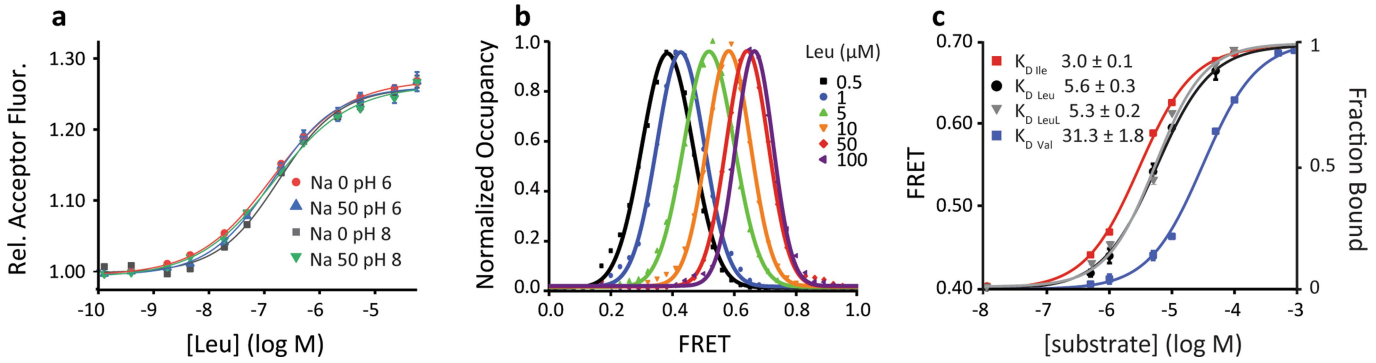
Competing interests S.C.B. has an equity interest in Lumidyne Technologies.

Additional information

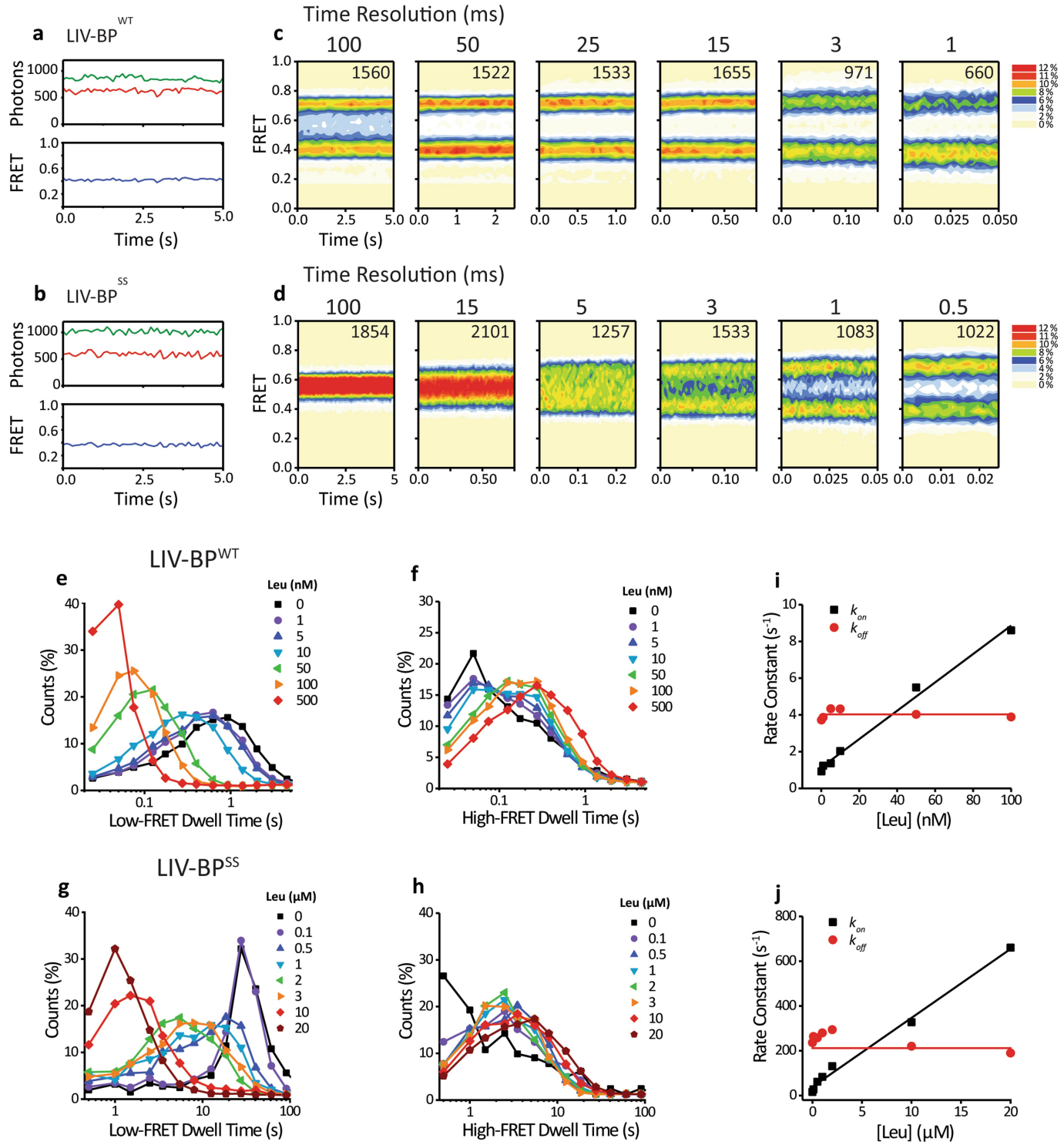
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1747-5.

Correspondence and requests for materials should be addressed to M.Q., J.A.J. or S.C.B.

Peer review information *Nature* thanks Antoine van Oijen, Robert Vandenberg and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

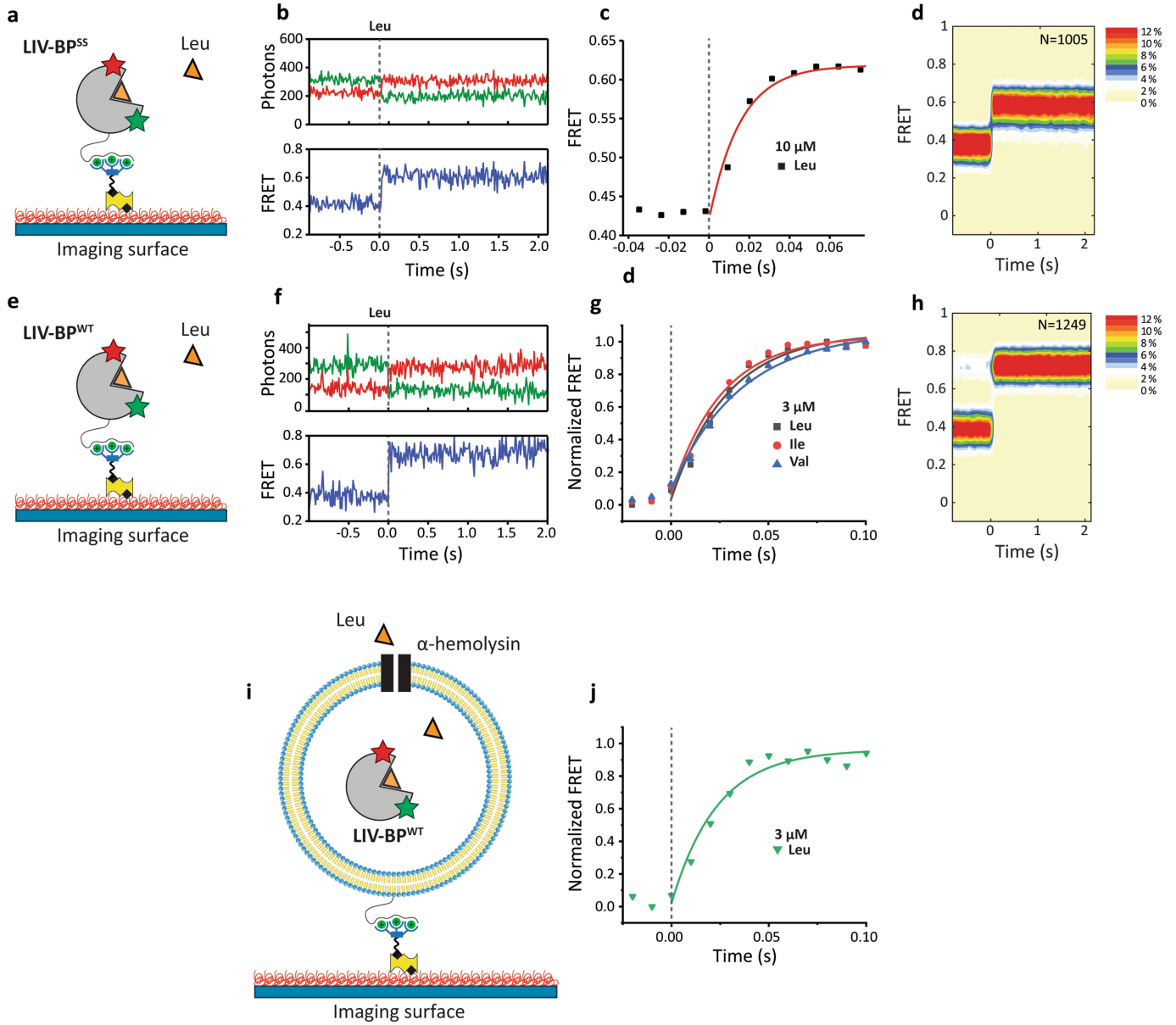


Extended Data Fig. 1 | Salt and pH dependence of LIV-BP^{WT} affinity for leucine and single-molecule calibration curve of LIV-BP^{SS}. a, Bulk acceptor fluorescence of LIV-BP^{WT} measured in the presence of varying leucine concentrations and buffer conditions, demonstrating a lack of dependence on salt and pH conditions. **b**, Surface-immobilized LIV-BP^{SS} incubated in the presence of varying concentrations of leucine. For each concentration, FRET efficiencies were summed across time to generate a histogram (symbols) and fit to Gaussian functions (lines) to obtain estimates of the mean FRET value. Data were collected at least three times with similar results. **c**, Mean FRET values for leucine (black circles), obtained as shown in a, fit with equation (1) in Supplementary Methods (black line) to obtain the K_d. Analogous data with mean FRET values are shown for each concentration for isoleucine (red), valine (blue) and leucine (black) were fit with equation (1) to determine the K_d. The grey data points and line (LeuL) indicate LIV-BP^{SS} encapsulated within the liposome with leucine allowed to equilibrate at each concentration for at least 10 min into the liposome lumen. This indicates that the binding activity of LIV-BP is unaffected by encapsulation within the liposome lumen. Mean ± s.e.m. (n = 3 separate days with 2 protein preparations).

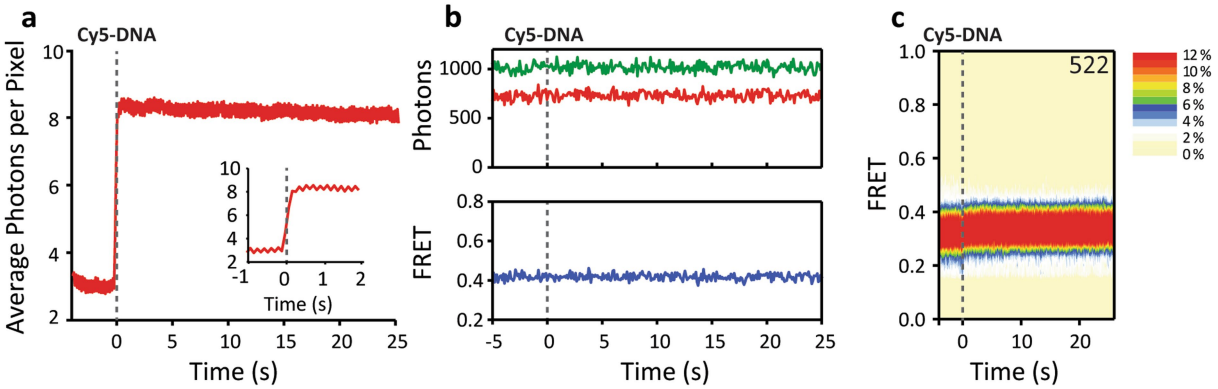


Extended Data Fig. 2 | See next page for caption.

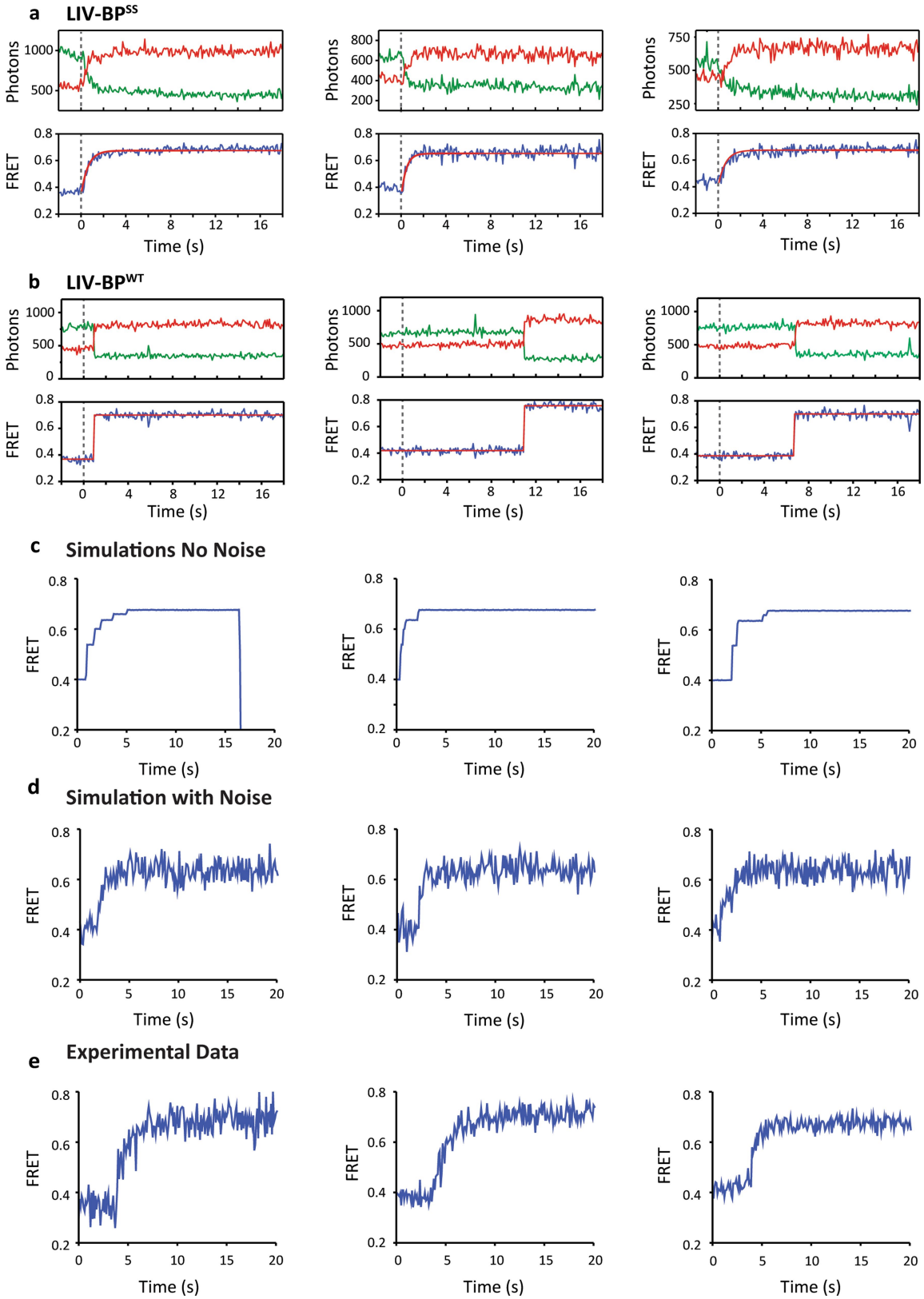
Extended Data Fig. 2 | Single-molecule imaging of LIV-BP^{WT} and LIV-BP^{SS} reveals kinetic determinants of analogue and digital responses. **a**, **b**, Representative traces of LIV-BP^{WT} (**a**) and LIV-BP^{SS} (**b**) imaged at 100-ms time resolution in the absence of substrate. **c**, **d**, LIV-BP^{WT} (**c**) and LIV-BP^{SS} (**d**) variants imaged in the presence of leucine at the K_d of each variant (40 nM and 5.6 μM, respectively) and at the indicated time resolution in milliseconds. FRET values from all selected traces (count at top right of each panel) summed into time-dependent population FRET histograms, represented as contour plots. Scale bar is shown at the right. Two distinct FRET peaks are apparent at all time resolutions for LIV-BP^{WT}; LIV-BP^{SS} displayed only a single peak at low time resolution (≥100 ms) that resolved into two district populations in the millisecond regime. **e**–**h**, LIV-BP^{WT} and LIV-BP^{SS} imaged at time resolutions that most-completely sampled the FRET transitions (25 ms and 0.25 ms, respectively) and idealized using the segmental *k*-means algorithm (Supplementary Methods). Dwell-time distributions of LIV-BP^{WT} in the low- (**e**) and high-FRET (**f**) states, and LIV-BP^{SS} in the low- (**g**) and high-FRET (**h**) states, in the presence of the indicated concentrations of leucine. All experiments were performed at least three times with similar results. **i**, **j**, Rate constants derived from maximum likelihood analysis of dwell times in the low- (black squares) and high-FRET (red circles) states, fit to lines to determine ligand association (k_{on}) (black lines) and ligand dissociation (k_{off}) (red lines) rate constants for LIV-BP^{WT} (**i**) and LIV-BP^{SS} (**j**). As expected for a bimolecular interaction, the ligand-binding rate increased linearly with ligand concentrations, whereas the ligand-dissociation rate remained constant. k_{on} values were similar in LIV-BP^{WT} and LIV-BP^{SS} at 77 and 30 $\mu M^{-1} s^{-1}$, respectively, whereas k_{off} values differed by nearly two orders of magnitude at 4.0 and 212 s^{-1}, respectively. Together, these results are consistent with the observed 100-fold difference in binding affinity in the two sensor variants (Fig. 1c). All experiments were performed at least three times with similar results.



Extended Data Fig. 3 | LIV-BP rapidly binds and responds to ligand. **a**, Schematic of His-tagged LIV-BP^{SS} directly surface-immobilized. **b**, Representative fluorescence (donor in green and acceptor in red) and FRET (blue) traces from a single LIV-BP^{SS} sensor imaged at 15-ms time resolution during the rapid delivery (vertical dashed line) of 10 µM (subsaturating) leucine. **c**, Ensemble-average FRET efficiency (symbols), fit to a single exponential function with a time-constant of approximately 16 ms (red line). **d**, FRET values of an ensemble of surface-immobilized LIV-BP^{SS} molecules summed into a contour plot (scale at right), demonstrating rapid and uniform response to leucine addition. **e**, Schematic of His-tagged LIV-BP^{WT} directly surface-immobilized. **f**, Representative fluorescence (donor in green and acceptor in red) and FRET (blue) traces from a single LIV-BP^{WT} sensor imaged at 10-ms time resolution during the rapid delivery (vertical dashed line) of 3 µM (saturating) leucine. **g**, Ensemble-average FRET efficiency of leucine (grey squares), isoleucine (red circles) or valine (blue triangles). Data are fit to a single exponential function with a time-constant of approximately 27 ms for leucine (grey line), 25 ms for isoleucine (red line) and 32 ms for valine (blue line). **h**, FRET histogram of an ensemble of surface-immobilized LIV-BP^{WT} molecules responding to leucine summed into a contour plot (scale at right), demonstrating rapid and uniform response to leucine addition. **i**, Schematic of LIV-BP^{WT} encapsulated within the lumen of liposomes in identical conditions to those used for transport assays (Supplementary Methods). Liposomes were pre-incubated in 100 µg ml^{-1} α-haemolysin for 15 min at room temperature. **j**, Encapsulated LIV-BP^{WT} FRET response to injection of 3 µM leucine, in which the time of mixing is marked (dashed grey line). The time constant of the observed FRET response (about 23 ms), fit as above (green line), was identical to LIV-BP^{WT} directly immobilized to the surface. All experiments were performed at least three times with similar results.

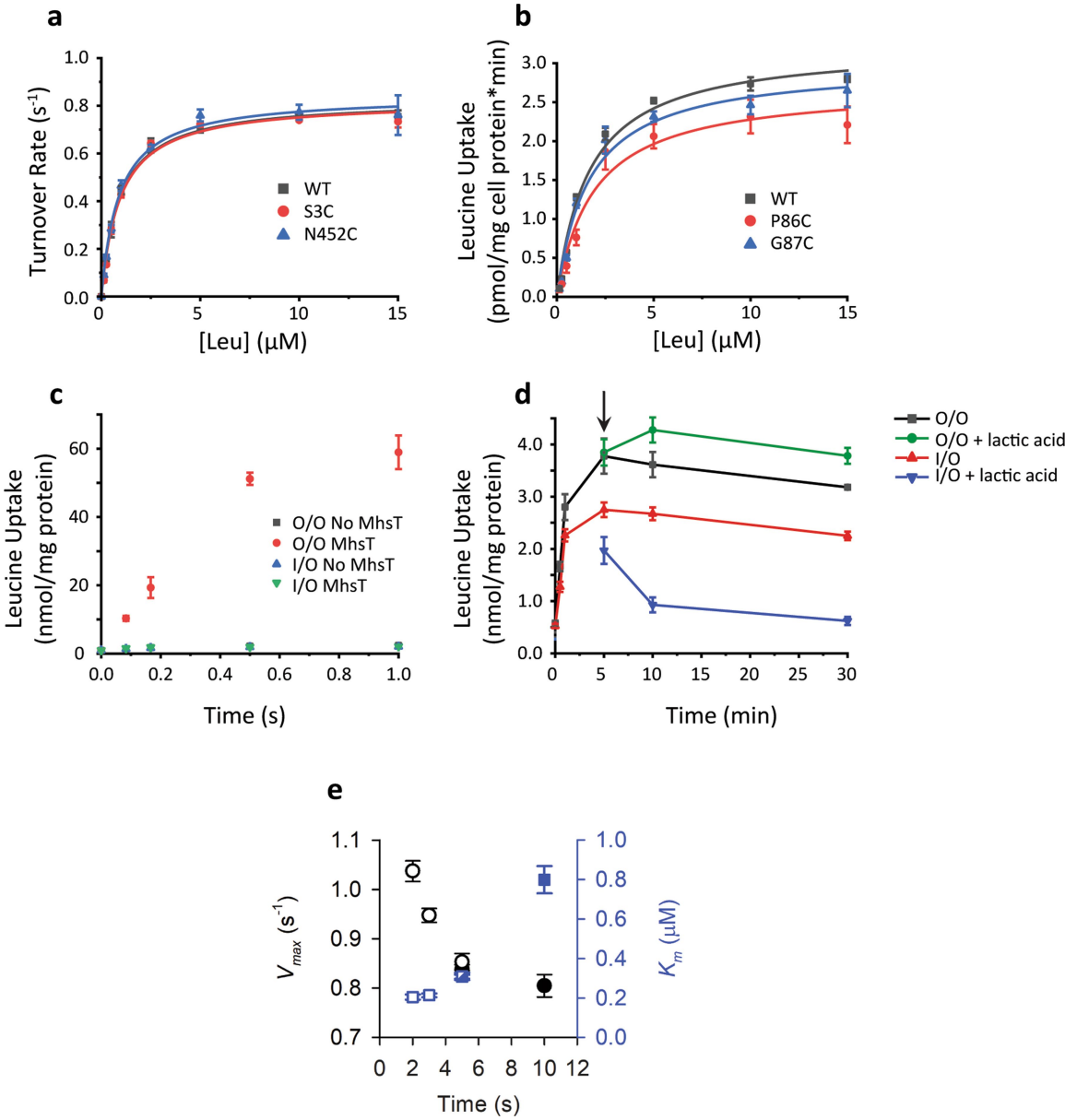


Extended Data Fig. 4 | Quantifying the timing of transport initiation. a, The precise timing of ligand injection was estimated by co-injecting a low concentration (0.5 nM) of Cy5-labelled 21-mer DNA duplex (Supplementary Methods) to the injected solution of interest, and measuring the increase in background fluorescence on the Cy5 channel in regions far from immobilized particles. The mean background fluorescence is shown, from a representative experiment in which LIV-BP^{WT}-containing liposomes that lack MhsT were imaged before and after (vertical dashed line) the injection of Cy5-labelled DNA tracer, with the exact time of injection determined by the midpoint of the step-like increase in fluorescence. Inset, zoomed-in view of the period immediately before and after injection. **b**, **c**, Representative single-molecule fluorescence (donor in green and acceptor in red) and FRET (blue) traces (**b**) and FRET contour plot (**c**) of encapsulated LIV-BP^{WT} from these experiments, demonstrating minimal changes in apparent FRET efficiency with co-injection of low concentrations of the Cy5 fluorophore. Experiments were performed at least three times with similar results.

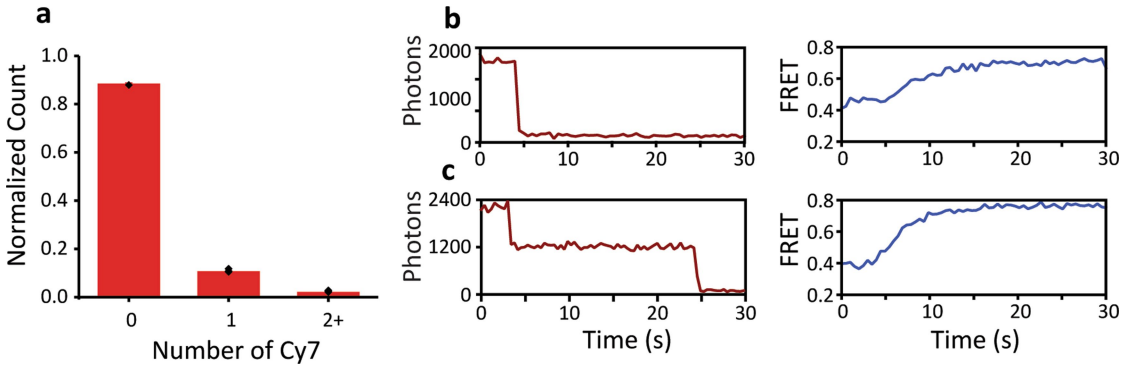


Extended Data Fig. 5 | See next page for caption.

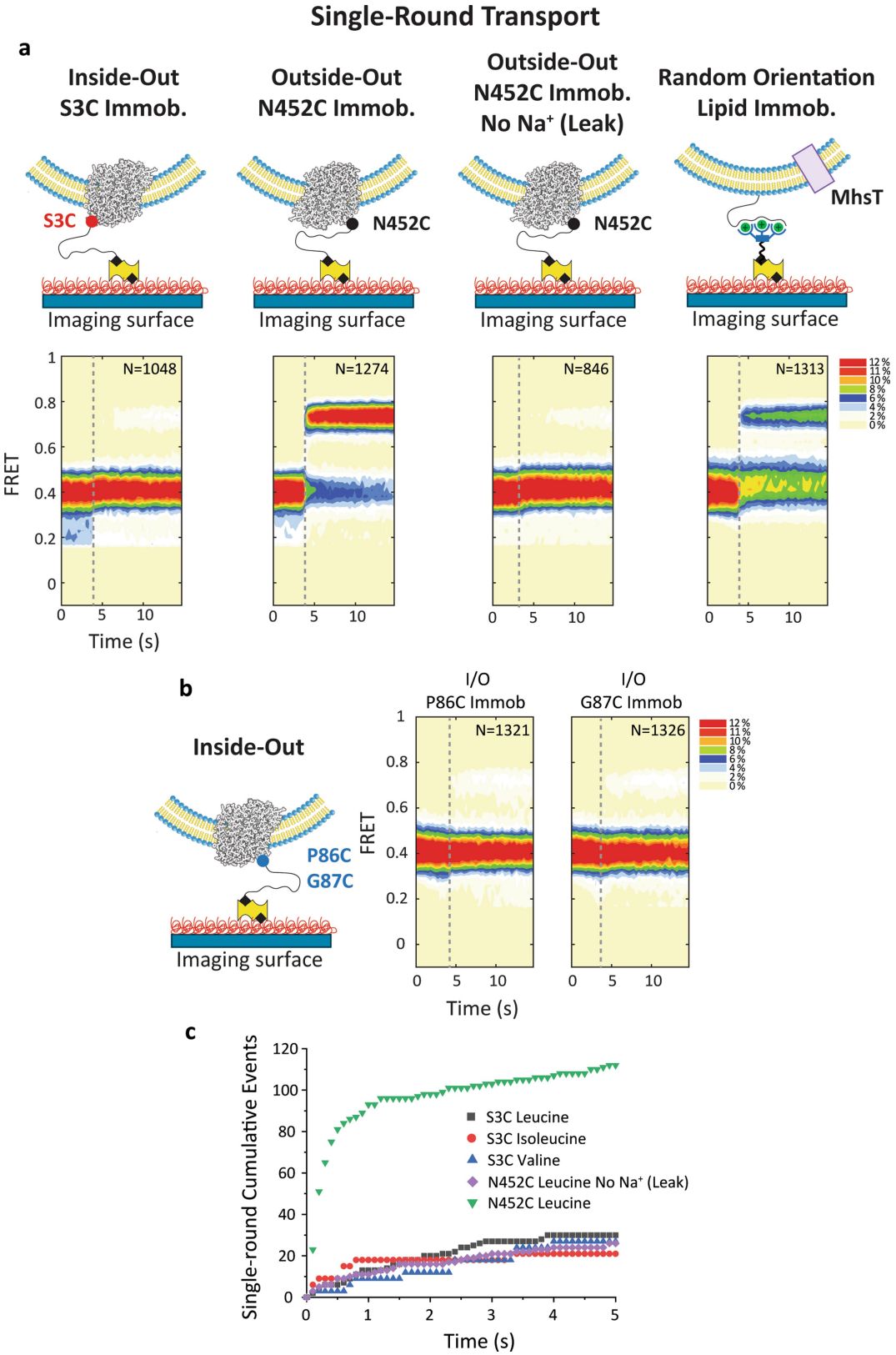
Extended Data Fig. 5 | Representative traces of multi-round and single turnover assays with simulations of multi-round activity. a, **b**, Single-molecule transport traces with 50 mM Na^+ and leucine (5 μM and 0.1 μM for a and **b**, respectively) at external pH 8 and internal pH 6, with the time of injection indicated by a dashed grey vertical line. **a**, Representative single-molecule fluorescence (top) (donor in green and acceptor in red) and FRET (bottom) (blue) traces and fits to exponential functions (red) from experiments imaging LIV-BP^{SS} encapsulated within proteoliposomes that contain MhsT. **b**, As in **a**, but for the single-turnover assay using LIV-BP^{WT} and state assignments in red (bottom panels). Traces shown are representative of experiments performed at least three times. **c**, Representative simulated FRET traces generated by a model in which distinct states correspond to the distinct FRET values that would be reported by the sensor, using the calibration curve according to the number of substrate molecules in the liposome (Extended Data Fig. 1). We assume transport occurs at a rate of one event per second and is irreversible, so transitions to lower-FRET states (state 1 to state 0) are not allowed. **d**, Noise that mimics experimental noise was added to representative FRET traces (different individual traces to those shown in c), which masks the individual steps observable in a. This indicates that, while in ideal circumstances we should be able to monitor individual transport events in this assay, in practice such an analysis would be difficult. **e**, Experimental FRET traces with the same apparent transport rate as the simulated data demonstrate a notable likeness to the simulated traces. Over 1,000 traces were simulated with similar results and representative traces were taken from experiments repeated at least 3 times.



Extended Data Fig. 6 | ^{3}H-leucine uptake activity of wild-type MhsT, and immobilization mutants and membranes with and without MhsT. **a**, Turnover rates of purified and reconstituted wild-type MhsT (grey) and Cy7-labelled MhsT(S3C) (red) and MhsT(N452C) (blue) assessed at a series of external leucine concentrations. Leucine uptake by wild-type MhsT exhibited a K_m of 0.93 ± 0.08 μM and a V_{max} of 0.83 ± 0.02 substrate molecules per second. MhsT(S3C) labelled with a Cy7 fluorophore had a K_m of 0.94 ± 0.11 μM and a V_{max} of 0.82 ± 0.03 substrate molecules per second. MhsT(N452C) labelled with a Cy7 fluorophore had a K_m of 0.90 ± 0.08 μM and a V_{max} of 0.84 ± 0.02 substrate molecules per second. **b**, Leucine uptake by MQ614 cells expressing wild-type MhsT (grey), MhsT(P86C) or MhsT(G87C) in the presence of 150 mM Na^+ at pH 8.5. Mean ± s.e.m. (n = 3 experiments). **c**, Inside-out and outside-out vesicles prepared (Supplementary Methods) and assayed for uptake activity in the presence and absence of the MhsT transporter as indicated. Transport is observed only in the presence of MhsT in the outside-out orientation. **d**, Radioactive glycine uptake by the native CycA glycine:H^+ symporters in inside-out or outside-out prepared vesicles in the presence of an inwardly directed proton gradient. Both preparations show robust activity, indicating that both vesicle preparations are intact and contain functional transporters. Lactic acid was added (arrow) to vesicles during the glycine-uptake time course to establish a proton gradient relative to the vesicle orientation. This creates an inwardly directed proton gradient in outside-out vesicles, and the opposite in inside-out vesicles. We observe an increase in the rate of glycine transport in outside-out vesicles and a marked decrease in inside-out vesicles, as expected, which demonstrates that we have prepared the vesicles in the indicated orientation. Mean ± s.e.m. (n = 3 experiments). **e**, The V_{max} and K_m of leucine uptake by wild-type MhsT were measured at a series of external leucine concentrations for the indicated periods of time. Assays were performed with proteoliposomes that contain known amounts of MhsT prepared at protein-to-lipid reconstitution ratios of 1:150 (w/w) (solid symbols) or 1:300 (w/w) (open symbols) for time periods of 2, 3, 5 or 10 s (total decays per minute at the corresponding time points were background-corrected for decays per minute determined at 0 s). The partially filled square indicates the virtual overlap of data points. The specific radioactivity–decays per minute correlation was verified using known amounts of ^{3}H-leucine. Shorter sampling times yielded higher turnover rates and lower K_m values (the highest V_{max} of 1.04 ± 0.02 s^{-1} and lowest K_m of 0.21 ± 0.01 μM were determined at the 2-s sampling time), but the technically challenging nature of these experiments precluded further shortening of the sampling time. To ensure the reliable determination of the V_{max} and K_m in radiolabelled uptake measurements, a sampling time of 3 s was chosen. Mean ± s.e.m. (n = 3 replicates of 2 protein preparations).

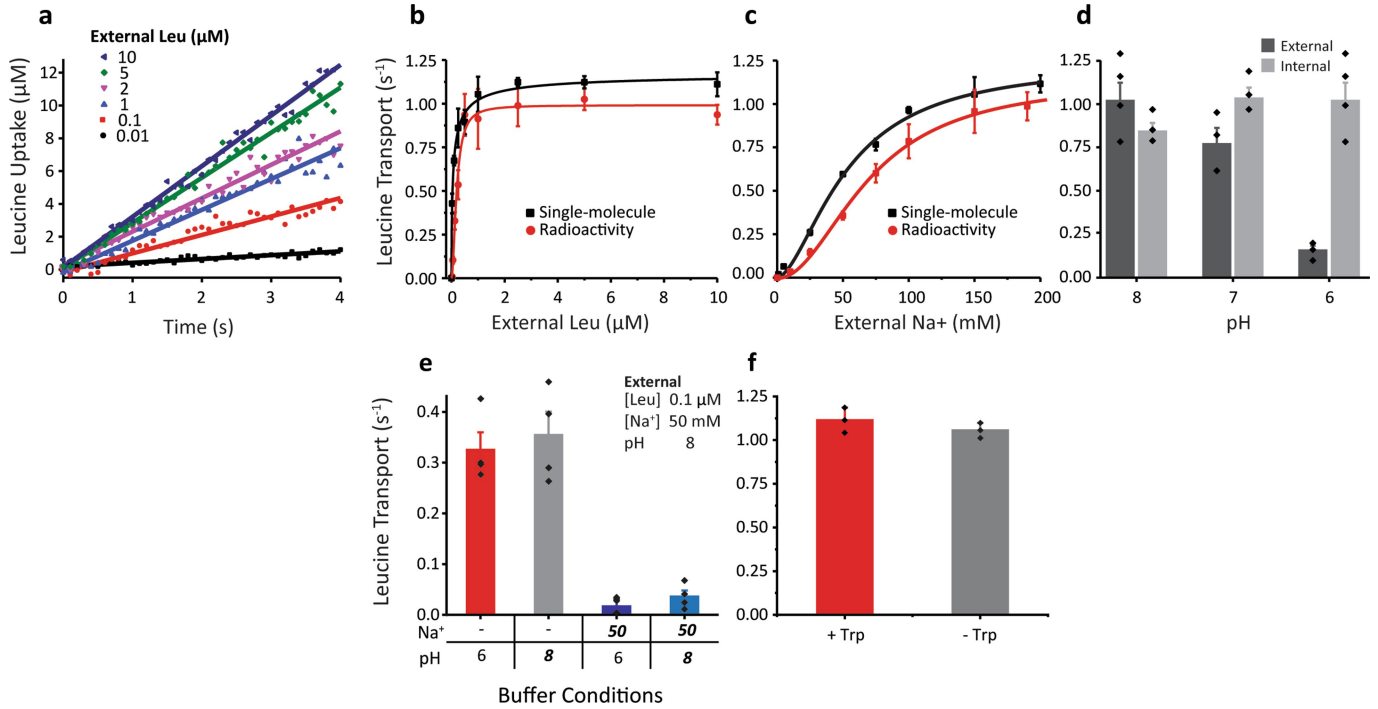


Extended Data Fig. 7 | Determination of the number of transporters per liposome. a, The number of Cy7-labelled transporters observed in each liposome. Reported values were corrected for a labelling efficiency of 75% to determine the final estimate of the number of transporters per liposome. Black symbols represent individual data points; bars represent mean ± s.e.m. ($n = 3$ experiments). **b**, **c**, Representative Cy7 fluorescence (left) and FRET (right) traces of a liposome with a single transporter (**b**) or two transporters (**c**). Traces are representatives of experiments performed at least three times.

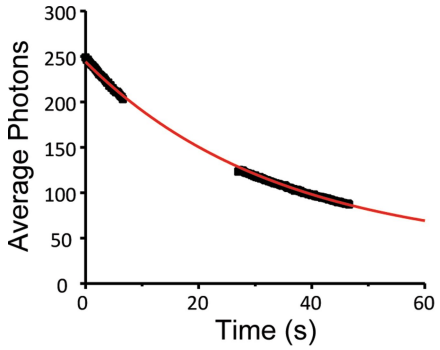


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Orientation-controlled single-turnover transport. **a**, Schematics and corresponding contour plots of the single-turnover assay with liposomes immobilized by the S3C (inside-out) or N452C (outside-out) residues of MhsT. The third panel from the left shows the passive diffusion of leucine in the absence of Na^+. The right panel shows transport data from proteoliposomes that contain a single (mixed orientation) MhsT transporter immobilized by His-tagged lipids. Occupancy of the high-FRET state following injection of leucine (grey dashed line) represents the proportion of vesicles in which transport has occurred. The N in top right corner indicates the total number of traces recorded over three separate experiments. **b**, Left, schematic of proteoliposome immobilization by biotin tags added at position P86C or G87C of MhsT, which result in inside-out orientations of the MhsT transporter (as for S3C). Right, contour plots of the single-turnover assay when immobilizing via biotin–P86C or biotin–G87C. The N in top right corner indicates the total number of traces over three separate experiments. **c**, Cumulative distributions of representative movies demonstrating the low translocation rates of the S3C (inside-out) immobilized liposomes with all three substrates, which match the leak rate in the absence of Na^+. Experiments were repeated at least three times with similar results.



Extended Data Fig. 9 | Michaelis–Menten kinetic parameters of MhsT transport. a, FRET values were transformed into units of proteoliposome luminal leucine concentration (Supplementary Methods) and fit to linear functions (lines) to determine transport initial rates, from experiments in the presence of the indicated concentrations of external leucine. Unless otherwise stated, experiments were performed with external 150 mM Na^+, 1 μM leucine and pH 8 outside with pH 6 inside, and with choline chloride used to maintain osmotic balance. **b**, Substrate accumulation rates from **a** (black squares) were fit to a Michaelis–Menten function (black line) with a K_m of 0.06 ± 0.01 μM and a V_{max} of 1.18 ± 0.04 s^{-1} with a Hill coefficient of 0.69 ± 0.08. Analogous radioactive uptake experiment (red circles) fit with a Michaelis–Menten function (red line) with a K_m of 0.19 ± 0.02 μM and a V_{max} of 0.99 ± 0.03 s^{-1} with a Hill coefficient of 1.85 ± 0.28. **c**, As in **b**, but varying the external concentration of Na^+. Black line is a fit to a Hill equation with a K_m of 69.9 ± 4.25 mM and a V_{max} of 1.14 ± 0.05 s^{-1} with a Hill coefficient of 2.03 ± 0.17. Analogous radioactive uptake experiment (red circles) fit with a Hill equation with a K_m of 63 ± 15 mM, V_{max} of 0.74 ± 0.11 μM s^{-1} and Hill coefficient of 1.61 ± 0.40. **d**, Substrate accumulation rates in the presence of the indicated external pH (dark grey, pH 6 inside) and internal pH (light grey, pH 8 outside) and 150 mM Na^+ and 1 μM leucine. The dark grey pH 8 bar and the light grey pH 6 bars are duplicated for comparison purposes. **e**, Substrate accumulation rates with 0.1 μM leucine pH 8 and 50 mM Na^+ external solution and in the presence and absence of 50 mM Na^+ at pH 6 and pH 8 of the internal solution. Robust transport is observed at both internal pH 6 and pH 8 in the absence of internal Na^+ (red). Addition of internal Na^+ completely abolishes transport (purple), with changes in internal pH having little to no further effect (cyan). Mean ± s.e.m. (n = 3+ experiments). **f**, Multi-round transport assay with 150 mM external Na^+, 1 μM leucine and pH 8 with (red) and without (black) internal 50 μM tryptophan at pH 6. There is no significant difference between the two conditions when tested using a two-sided Student's t-test that was not corrected for multiple hypotheses at a 95% confidence interval. Black diamonds represent individual data points, and bars represent mean ± s.e.m. (n = 3 experiments).



Extended Data Fig. 10 | Proton leak into proteoliposomes monitored by lipid-linked pHrodo fluorophores. Buffer was exchanged from pH 6 to pH 8 (outside) and a step-like approximately 50% decrease in fluorescence was observed, which we interpret as the quenching of lipid-linked pHrodo on the liposome exterior, followed by a slow decay in fluorescence (black squares). The data were fit with a single exponential function (red line) with a time constant of about 33 s, which we interpret as the slow leak of protons into the liposome. The gaps in the data are periods in the absence of illumination, to rule out photobleaching as a cause of the fluorescence decay we observed. These experiments confirm that the pH gradient was maintained during the full time-frame in which we measure transport. Experiments were performed three times with similar results.

nature research

Corresponding author(s): Dr. Scott Blanchard

Last updated by author(s): Sep 23, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Commercial products were used (Labview 2017)
Data analysis	All software are commercial products (OriginPro 8, ImageJ 1.4, and MATLAB R2018b) or otherwise publicly available (SPARTAN).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes of 500-1,000 molecules per condition/repeat were determined by the normal throughput of the instrumentation and were found to adequately sample the distribution of behaviors in the ensemble. Statistical methods were not used to determine sample sizes. Standard replicate sizes in the field were used (n=3+)
Data exclusions	Exclusion criteria to remove clear artifacts in single molecule traces were predefined and uniformly applied, as described in the manuscript.
Replication	All findings were reliably replicated on separate days with fresh buffer solutions and frozen sample aliquots from multiple preparations. All results shown were successfully replicated.
Randomization	Samples were not randomized as the experimental procedure is systematic, the selection criteria were uniform across experiments and data analysis was largely automated and therefore identical across replicates. Additionally, the large number of total replicated performed in the study (300+) would have made randomization infeasible.
Blinding	Blinding was not performed as the experimental procedure is systematic, the selection criteria were uniform across experiments and data analysis was largely automated and therefore identical across replicates. Additionally, the large number of total replicated performed in the study (300+) would have made randomization infeasible.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Article

Cryo-EM structure of the spinach cytochrome b_6f complex at 3.6 Å resolution

https://doi.org/10.1038/s41586-019-1746-6

Received: 16 May 2019

Accepted: 8 October 2019

Published online: 13 November 2019

Lorna A. Malone[1], Pu Qian[1], Guy E. Mayneord[1], Andrew Hitchcock[1], David A. Farmer[1], Rebecca F. Thompson[2], David J. K. Swainsbury[1], Neil A. Ranson[2], C. Neil Hunter[1,3*] & Matthew P. Johnson[1,3*]

The cytochrome b_6f (cytb_6f) complex has a central role in oxygenic photosynthesis, linking electron transfer between photosystems I and II and converting solar energy into a transmembrane proton gradient for ATP synthesis[1–3]. Electron transfer within cytb_6f occurs via the quinol (Q) cycle, which catalyses the oxidation of plastoquinol (PQH_2) and the reduction of both plastocyanin (PC) and plastoquinone (PQ) at two separate sites via electron bifurcation[2]. In higher plants, cytb_6f also acts as a redox-sensing hub, pivotal to the regulation of light harvesting and cyclic electron transfer that protect against metabolic and environmental stresses[3]. Here we present a 3.6 Å resolution cryo-electron microscopy (cryo-EM) structure of the dimeric cytb_6f complex from spinach, which reveals the structural basis for operation of the Q cycle and its redox-sensing function. The complex contains up to three natively bound PQ molecules. The first, PQ1, is located in one cytb_6f monomer near the PQ oxidation site (Q_p) adjacent to haem b_p and chlorophyll *a*. Two conformations of the chlorophyll *a* phytyl tail were resolved, one that prevents access to the Q_p site and another that permits it, supporting a gating function for the chlorophyll *a* involved in redox sensing. PQ2 straddles the intermonomer cavity, partially obstructing the PQ reduction site (Q_n) on the PQ1 side and committing the electron transfer network to turnover at the occupied Q_n site in the neighbouring monomer. A conformational switch involving the haem c_n propionate promotes two-electron, two-proton reduction at the Q_n site and avoids formation of the reactive intermediate semiquinone. The location of a tentatively assigned third PQ molecule is consistent with a transition between the Q_p and Q_n sites in opposite monomers during the Q cycle. The spinach cytb_6f structure therefore provides new insights into how the complex fulfils its catalytic and regulatory roles in photosynthesis.

Photosynthesis sustains life on Earth by converting light into chemical energy in the form of ATP and NADPH, producing oxygen as a by-product. Two light-powered electron transfer reactions at photosystems I and II (PSI and PSII) are linked via the cytb_6f complex to form the 'Z-scheme' of photosynthetic linear electron transfer (LET)[1]. Cytb_6f catalyses the rate-limiting step in LET, coupling the oxidation of PQH_2 and reduction of PC and PQ to the generation of a transmembrane proton gradient, which is used by ATP synthase to make ATP[2,3]. The cytb_6f complex is analogous to the cytbc_1 complex found in mitochondria[4] and anoxygenic photosynthetic bacteria[5], and both operate via the modified Q cycle[2,6]. The cytb_6f and cytbc_1 complexes are dimeric and have similarly arranged electron transfer cofactors, comprising a 2Fe-2S cluster, two *b*-type haems and a *c*-type haem. However, crystallographic structures of cyanobacterial and algal cytb_6f complexes have revealed additional cofactors that are not found in cytbc_1 complexes, including chlorophyll *a*, 9-*cis* β-carotene and an additional *c*-type high-spin haem[7–9]. The Q-cycle involves bifurcated transfer of two electrons, derived from oxidizing a lipophilic PQH_2 molecule at the Q_p-binding site, into the high- (2Fe-2S, cyt*f*) and low- (cytb_p, b_n and c_n) redox potential pathways, whereas the two protons enter the thylakoid lumen[2,6]. The high-potential pathway delivers an electron to a membrane-extrinsic soluble acceptor protein, PC, destined for PSI, while the low potential pathway delivers its electron to a PQ molecule bound at the Q_n site near the stromal side of the membrane. Oxidation of a second PQH_2 at the Q_p site culminates in the two-electron reduction of a Q_n site bound PQ, which together with two proton transfers from the stroma, regenerates PQH_2. The Q-cycle thereby doubles the number of protons transferred to the lumen per PQH_2 oxidized[2,6]. Yet, full understanding of the Q-cycle mechanism is hindered by a lack of information on the binding of the substrate PQ/ PQH_2 molecules within the complex.

In addition to its role in LET, cytb_6f also plays a key part as a redox sensing hub involved in the regulation of light harvesting and cyclic

[1]Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield, UK. [2]Astbury Centre for Structural Molecular Biology, School of Molecular & Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds, UK. [3]These authors jointly supervised this work: C. Neil Hunter, Matthew P. Johnson. *e-mail: c.n.hunter@sheffield.ac.uk; matt.johnson@sheffield.ac.uk

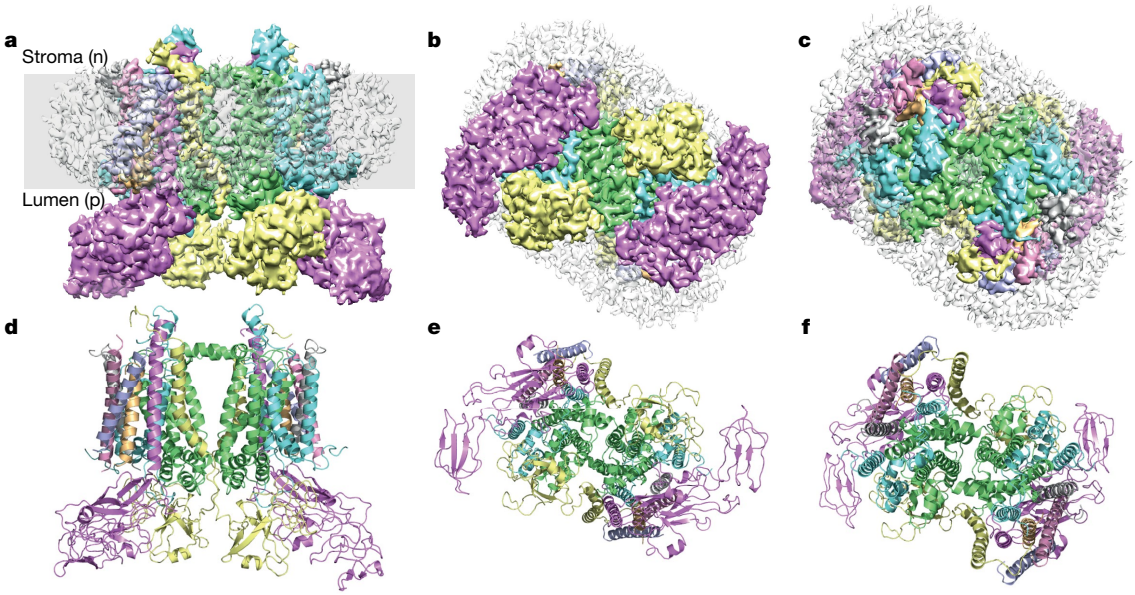


Fig. 1 | Cryo-EM structure of the cytb_6f complex from spinach. a–c, Views of the colour-coded cytb_6f density map showing cytb_6 (green), cyt$f$ (magenta), ISP (yellow), subunit IV (cyan), PetG (grey), PetM (pink), PetN (pale orange) and PetL (pale purple). Detergent and other disordered molecules are shown in semi-transparent light grey. **a**, View in the plane of the membrane. The grey stripe indicates the probable position of the thylakoid membrane bilayer. **b**, View perpendicular to the membrane plane from the lumenal (p) side. **c**, View perpendicular to the membrane plane from the stromal (n) side. **d–f**, Modelled subunits of cytb_6f shown in a cartoon representation and coloured as in **a–c**. **d**, View in the plane of the membrane. **e**, View perpendicular to the membrane plane from the lumenal side. **f**, View perpendicular to the membrane plane from the stromal side.

electron transfer (CET), which optimize photosynthesis in fluctuating light environments[10,11]. Cytb_6f communicates the redox status of the PQ pool to the loosely associated light harvesting complex II (LHCII) kinase, STN7[12–14]. Phosphorylation of LHCII results in a decrease in thylakoid membrane stacking, promoting the exchange of LHCII between PSII and PSI to balance their relative excitation rates[15] and regulate CET[16]. CET involves the reinjection of electrons from ferredoxin into the PQ pool, generating a proton gradient for photoprotective downregulation of PSI and PSII or to augment ATP synthesis, without net formation of NADPH[11]. The cytb_6f complex has been proposed to fulfil the role of the ferredoxin–PQ oxidoreductase (FQR) during CET, with the stromal-facing haem c_n suggested to channel electrons from ferredoxin NADP$^+$ reductase (FNR) bound ferredoxin to the Q_n-site PQ[17]. How cytb_6f performs these central redox-sensing regulatory roles remains unclear.

Genetic manipulation of photosynthetic regulation is now recognized as being key to increasing crop yields to feed a global population projected to approach 10 billion by 2050[18]. Indeed, overproduction of the Rieske iron–sulfur protein (ISP) of cytb_6f in *Arabidopsis thaliana* led to a 51% increase in yield[19]. Further progress in understanding the regulatory roles of cytb_6f and potentially manipulating them for crop improvement requires knowledge of the structure of the higher plant complex. Here, using a gentle purification procedure to obtain a highly active dimeric complex (Extended Data Fig. 1) and single-particle cryo-EM, we resolve the cytb_6f complex from *Spinacia oleracea* (spinach) at 3.6 Å resolution (Extended Data Fig. 2, Extended Data Table 1).

The colour-coded map (Fig. 1a–c) shows the architecture of this dimeric complex surrounded by a disordered density comprising detergent and lipid molecules. The overall organization of this higher plant cytb_6f complex is similar to crystallographic structures of algal and cyanobacterial complexes from *Chlamydomonas reinhardtii*[7] (Protein Data Bank (PDB): 1Q90), *Mastigocladus laminosus*[8] (PDB: 1VF5) and *Nostoc* sp. PCC 7120[9] (PDB: 2ZT9) (Extended Data Table 2). Each monomeric unit of the cytb_6f complex comprises four large polypeptide subunits that contain redox co-factors (cyt$f$, cytb_6, ISP and subunit IV), and four small subunits (PetG, PetL, PetM and PetN). Extended Data Figure 3 shows the density and structural model for each subunit. The extrinsic domains of cyt$f$ and the ISP on the lumenal face of the complex flank the membrane-integral cytb_6 subunits (Fig. 1a, b). The organization of the transmembrane integral subunits can be seen on the stromal side of the complex (Fig. 1c), with 13 transmembrane helices visible within each monomer (Fig. 1d–f). Peripheral to the core of cytb_6 (four transmembrane helices) and subunit IV (three transmembrane helices) on the long axis of the complex is the single kinked transmembrane helix of the ISP that crosses over to provide the soluble ISP domain of the neighbouring monomer. The single transmembrane helix belonging to cyt$f$ is sandwiched by the transmembrane helices of the four minor subunits PetG, PetL, PetM and PetN, which form a 'picket fence' at the edge of the complex.

Figure 2a, b shows the organization of the prosthetic groups and lipids, with four c-type haems (f and c_n, dark blue), four b-type haems (b_p and b_n, red), two 2Fe-2S clusters (burnt orange and yellow), two 9-*cis* β-carotenes (orange), two chlorophyll a molecules (green), three PQ molecules (yellow) and twelve bound lipids (two monogalactosyl diacylglycerol, four phosphatidylglycerol, three sulfoquinovosyl diacylglycerol and three phosphatidylcholine, all shown in white). Extended Data Figure 4 shows the density map and structural model for each prosthetic group. Figure 2c shows all of the bound electron transfer cofactor edge-to-edge distances within the cytb_6f complex. Electron transfer from the 2Fe-2S cluster is thought to involve movement of the lumenal ISP domain, pivoting between closer association with the Q_p site and the haem f. In comparison to the chicken cytbc_1 complex, in which the two conformations of the ISP were resolved[20], the ISP and bound 2Fe-2S cluster in the spinach cytb_6f structure appear to be in the distal position with respect to haem f, as in the existing algal and cyanobacterial cytb_6f structures (Extended Data Table 2). PQ locations are generally inferred from crystallographic structures containing tightly bound quinone analogue inhibitors[21–23]. The cryo-EM structure was obtained with native PQ molecules (Fig. 2d), clearly defined by their respective densities (Extended Data Fig. 4); their distances from the nearest cofactors are shown in Fig. 2e–g. One PQ molecule (PQ1) is adjacent to the haem b_p and chlorophyll on one side of the dimer (Fig. 2e), and a second (PQ2) binds adjacent to the haem c_n–haem b_n

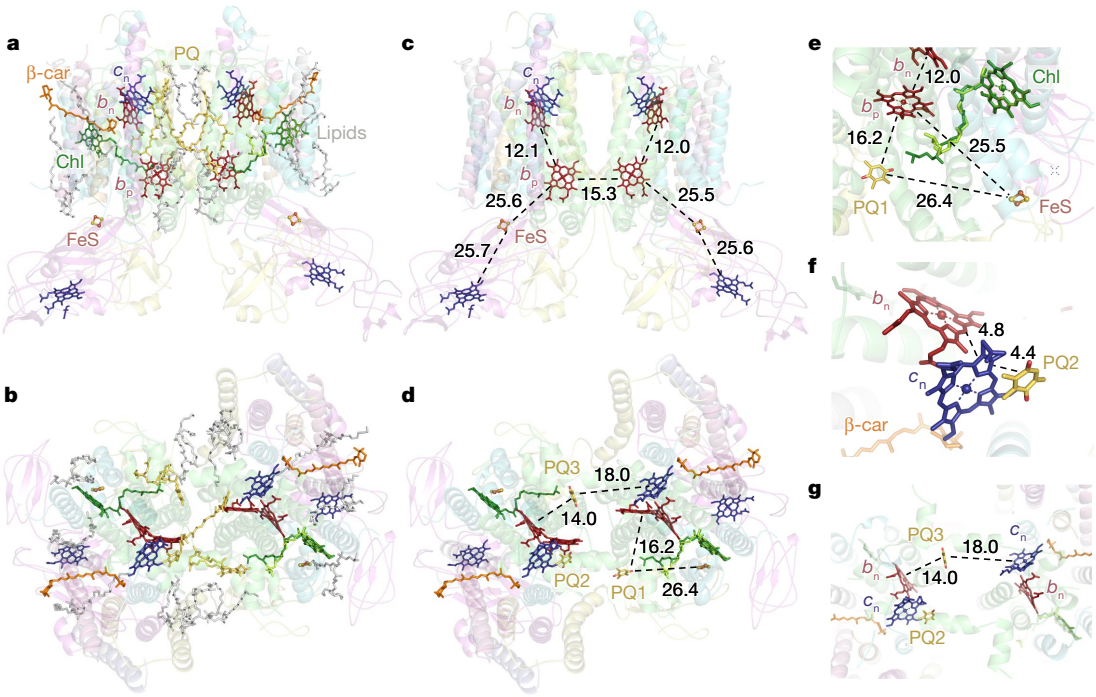


Fig. 2 | The global arrangement of prosthetic groups, lipids and plastoquinone molecules in the spinach cytb_6f complex. a, **b**, The arrangement of molecules in the cytb_6f complex viewed in the membrane plane (**a**) and perpendicular to the membrane plane from the stromal side (**b**). Chl, chlorophyll *a*; b_n, haem b_n; c_n, haem c_n; b_p, haem b_p; β-car, 9-*cis* β-carotene; FeS, 2Fe-2S. **c**, **d**, Cofactors and edge-to-edge distances (in Å) in the dimeric cytb_6f complex. **e**, The location of the 1,4-benzoquinone ring of PQ1 adjacent to haem b_p, the 2Fe-2S centre and two conformations of the chlorophyll molecule, represented in two shades of green. **f**, Close-up of the 1,4-benzoquinone ring of PQ2 and the nearby haem c_n and haem b_n near the stromal face of the complex. **g**, The 1,4-benzoquinone ring of PQ3, which sits between the haem c_n and haem b_n from the two cytb_6f monomers. The cytb_6f complex is coloured as in Fig. 1, and shows *c*-type haems (*f* and c_n, dark blue), *b*-type haems (b_p and b_n, red), 9-*cis*-β-carotene (orange), chlorophyll *a* (green), 2Fe-2S (burnt orange and yellow), lipids (white) and plastoquinones (PQ1–PQ3; yellow).

pair on the opposite monomer to PQ1 (Fig. 2f). A third and less clearly defined PQ (PQ3) lies between the haem c_n of one monomer and the haem b_n of the other (Fig. 2g). The density map in this region could also be interpreted as a phospholipid; Extended Data Fig. 6 shows the two possible fits—to a plastoquinone or a lipid.

The 1,4-benzoquinone ring of PQ1 is 16.2 Å from haem b_p and 26.4 Å from the 2Fe-2S cluster (Fig. 2e), and distal to the Q_p quinone oxidizing site defined in the *M. laminosus* cytb_6f structure[23] (PDB: 4H13) by the inhibitor tridecylstigmatellin (Fig. 3a, b). The Q_p site is located in a pocket formed by hydrophobic residues from subunit IV (Val84, Leu88, Val98 and Met101) and cytb_6 (Phe81, Val126, Ala129, Val133, Val151 and Val154) (Fig. 3c). Bifurcated electron transfer to the 2Fe-2S cluster and haem b_p involves two deprotonation events mediated by the ISP His128 and subunit IV Glu78 residues[2,3], which are buried inside the Q_p pocket (Fig. 3a, b). The OH group of PQ1 is ~26 Å from His128, a ligand of the 2Fe-2S cluster (Fig. 3b), so PQ1 is unlikely to be oxidized in its resolved position, which probably represents a snapshot of its approach to the Q_p site. It is notable in this regard that our spinach cytb_6f structure resolves two conformations of the chlorophyll phytyl tail, one of which permits access to Q_p site and one that restricts it (Fig. 3c, d). There is

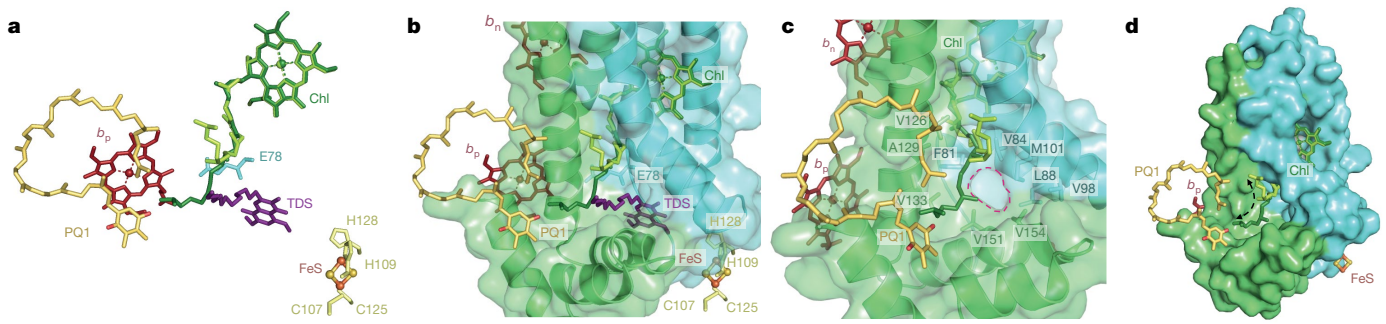


Fig. 3 | Conformational alterations in the chlorophyll phytyl chain at the PQH_2-oxidizing Q_p site. a, Orientation of the PQ1 in relation to the haem b_p, chlorophyll and 2Fe-2S cofactors. The catalytically essential residue E78 and coordinating residues of the 2Fe-2S cofactor are shown. Tridecylstigmatellin (TDS) is a quinone analogue, superimposed according to its position determined in the cyanobacterial complex (PDB: 4H13)[23], and used here to indicate the probable destination of PQ1 in the Q_p pocket. **b**, The same cofactors and residues as in **a**, but in relation to a surface view of cytb_6 (green) and subunit IV (sub IV, cyan). **c**, The Q_p pocket is highlighted with a red dashed line, showing its position in relation to the chlorophyll and PQ1 molecules; the hydrophobic residues of subunit IV (cyan) and cytb_6 that line the pocket are shown as sticks and coloured cyan and green, respectively. **d**, The two conformations of the chlorophyll tail (represented in dark green and light green) gate (dashed arrow) the entrance to the Q_p pocket.

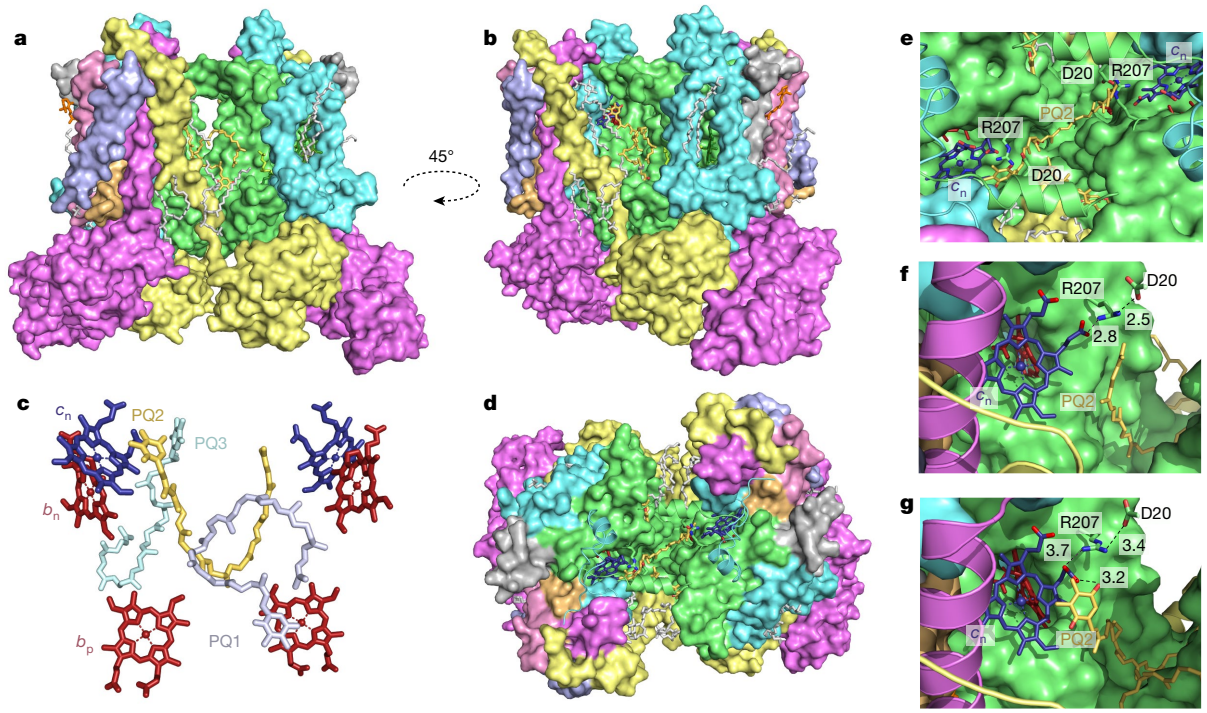


Fig. 4 | The intermonomer cavity of the spinach cytb_6f complex. a, **b**, Surface representations of the complex, with subunits coloured as in Fig. 1, and cofactors and lipids coloured as in Fig. 2. These two views of the complex are related by a 45° rotation about an axis perpendicular to the membrane to show two views of the cavity and the locations of PQ molecules. **c**, PQ1–PQ3 are shown in relation to the b_n, c_n and b_p haems in the core of the complex, viewed in the membrane plane. **d**, The complex viewed from the stromal side of the membrane; peripheral helices of cytb_6 and subunit IV are shown in cartoon representation for clarity, to show PQ2 straddling the intermonomer cavity and sitting between the two c_n haems. **e**, Close-up of the cavity in **d**. **f**, **g**, The head and tail regions of PQ2 in relation to the c_n haems on both sides of the cavity, highlighting the different orientations of the haem c_n propionates, and the Arg207 and Asp20 side chains. The distances in angstroms between the residues and cofactors are labelled.

only one position of the phytyl tail for the chlorophyll on the opposing monomer. The bound chlorophyll adjacent to PQ1 may fulfil a gating function at the Q_p pocket, either controlling access of PQH_2 and/or increasing the retention time of the reactive semiplastoquinone (SPQ) intermediate species formed following electron transfer to the 2Fe-2S cluster. Indeed, spin-coupling between the SPQ and the 2Fe-2S cluster has been detected during enzymatic turnover of cytb_6f but is absent in cytbc_1 complexes that lack the chlorophyll molecule[24]. SPQ in the 2Fe-2S-bound state does not react with oxygen, providing a potential mechanism to control the release of superoxide from the Q_p site[24] and regulate the activity of the LHCII kinase STN7[25], which is proposed to bind to cytb_6f between transmembrane helices F and H of subunit IV[26]. Another role for chlorophyll in regulating the activity of STN7 could involve PQH_2 displacing the chlorophyll phytyl tail on binding to the Q_p site; this motion could induce a conformational change in STN7 leading to its activation[27].

PQ2 binds towards the stromal face of the complex, 4.4 Å from the haem c_n–b_n pair at the Q_n reducing site (Fig. 2f). The b_n and c_n haems on each monomer are separated by 4.9 Å, with the b_n haem coordinated by His202 and His100 (cytb_6), whereas the vinyl side-chain of haem c_n is covalently linked to Cys35 (cytb_6) (Fig. 2f). The dimerization interface of the cytb_6f complex creates a cavity, which is proposed to promote transfer of quinones between the Q_p and Q_n sites on neighbouring monomers[8] (Fig. 4a, b). It is noteworthy that the three resolved PQ molecules inhabit this cavity and that PQ2 assumes a position 'diagonally' opposite to PQ1 (Fig. 4c) on the other monomer, as previously suggested[28]. PQ2 adopts a bowed conformation that straddles the intermonomer cavity with the distal PQ2 tail appearing to partially obstruct the Q_n site in the neighbouring monomer (Fig. 4d, e). This arrangement may have functional importance in preventing the simultaneous binding of PQ molecules at both Q_n sites, avoiding competition for electrons and favouring faster turnover of the Q cycle. Rapid provision of two electrons for PQ2 bound at a particular Q_n site could be facilitated by the 15.3 Å electron-tunnelling distance between b_p haems (Fig. 2c), which enables rapid inter-monomer electron transfer via the 'bus-bar' model from the neighbouring low-potential chain[29,30]. Alternatively, the second electron could be provided to the haem c_n directly via an FNR–ferredoxin complex bound at the stromal surface via CET[17,28]. The haem c_n propionates on the two halves of the cytb_6f dimer adopt different conformations (Fig. 4f, g); in the PQ-vacant site on the opposing monomer, the haem c_n propionate is more closely associated with Arg207 (Fig. 4f), whereas in the PQ-occupied site, the haem c_n propionate is rotated towards the 1,4-benzoquinone ring of PQ2 (Fig. 4g). The altered ligation of haem c_n on PQ binding is consistent with the downshift of its redox potential[31], which would strongly favour PQ reduction. We note that the reduction and oxidation of haem c_n is accompanied by the binding and release of one proton[31] so only one proton is required from the stromal side via the Arg207 and Asp20 residues (Fig. 4f, g) for PQ2 reduction to proceed rapidly, avoiding SPQ formation. It is also possible to position an oppositely oriented PQ within the density map, albeit with a less satisfactory fit (Extended Data Fig. 5). A third PQ molecule (PQ3) (Fig. 2g) has been assigned to the density between the Q_p and Q_n binding sites (see Extended Data Fig. 6 for an alternative assignment as phosphatidylcholine) with the 1,4-benzoquinone ring near the channel that links the two sides of the intermonomer cavity and the isoprenyl tail at the mouth of the Q_p site. This third PQ may therefore capture a snapshot of the molecule transitioning between the Q_p and Q_n sites in opposite monomers.

The cryo-EM structure of spinach cytb_6f reveals the positions of natively bound PQ and provides details regarding the conformational switches involved in PQ binding to the Q_n site, chlorophyll gating of the Q_p site and PQ exchange between the sites during Q-cycle operation.

Online content

Any methods, additional references, Nature Research reporting summaries, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1746-6.

1. Hill, R. & Bendall, F. Function of the two cytochrome components in chloroplasts: a working hypothesis. *Nature* **186**, 136–137 (1960).
2. Cramer, W. A., Hasan, S. S. & Yamashita, E. The Q cycle of cytochrome *bc* complexes: a structure perspective. *Biochim. Biophys. Acta* **1807**, 788–802 (2011).
3. Tikhonov, A. N. The cytochrome b_6f complex at the crossroad of photosynthetic electron transport pathways. *Plant Physiol. Biochem.* **81**, 163–183 (2014).
4. Xia, D. et al. Crystal structure of the cytochrome bc_1 complex from bovine heart mitochondria. *Science* **277**, 60–66 (1997).
5. Esser, L. et al. Inhibitor-complexed structures of the cytochrome bc_1 from the photosynthetic bacterium *Rhodobacter sphaeroides*. *J. Biol. Chem.* **283**, 2846–2857 (2008).
6. Cape, J. L., Bowman, M. K. & Kramer, D. M. Understanding the cytochrome *bc* complexes by what they don't do. The Q-cycle at 30. *Trends Plant Sci.* **11**, 46–55 (2006).
7. Stroebel, D., Choquet, Y., Popot, J.-L. & Picot, D. An atypical haem in the cytochrome b_6f complex. *Nature* **426**, 413–418 (2003).
8. Kurisu, G., Zhang, H., Smith, J. L. & Cramer, W. A. Structure of the cytochrome b_6f complex of oxygenic photosynthesis: tuning the cavity. *Science* **302**, 1009–1014 (2003).
9. Baniulis, D. et al. Structure–function, stability, and chemical modification of the cyanobacterial cytochrome b_6f complex from *Nostoc* sp. PCC 7120. *J. Biol. Chem.* **284**, 9861–9869 (2009).
10. Bellafiore, S., Barneche, F., Peltier, G. & Rochaix, J. D. State transitions and light adaptation require chloroplast thylakoid protein kinase STN7. *Nature* **433**, 892–895 (2005).
11. Yamori, W. & Shikanai, T. Physiological functions of cyclic electron transport around photosystem I in sustaining photosynthesis and plant growth. *Annu. Rev. Plant Biol.* **67**, 81–106 (2016).
12. Horton, P. & Black, M. T. Activation of adenosine 5′-triphosphate induced quenching of chlorophyll fluorescence by reduced plastoquinone. The basis of state I–state II transitions in chloroplasts. *FEBS Lett.* **119**, 141–144 (1980).
13. Vener, A. V., van Kan, P. J. M., Rich, P. R., Ohad, I. & Andersson, B. Plastoquinol at the quinol oxidation site of reduced cytochrome *bf* mediates signal transduction between light and protein phosphorylation: thylakoid protein kinase deactivation by a single-turnover flash. *Proc. Natl Acad. Sci. USA* **94**, 1585–1590 (1997).
14. Gal, A., Hauska, G., Herrmann, R. & Ohad, I. Interaction between light harvesting chlorophyll-*a*/*b* protein (LHCII) kinase and cytochrome b_6/f complex. In vitro control of kinase activity. *J. Biol. Chem.* **265**, 19742–19749 (1990).
15. Allen, J. F., Bennett, J., Steinback, K. E. & Arntzen, C. J. Chloroplast protein phosphorylation couples plastoquinone redox state to distribution of excitation energy between photosystems. *Nature* **291**, 25–29 (1981).
16. Wood, W. H. J. et al. Dynamic thylakoid stacking regulates the balance between linear and cyclic photosynthetic electron transfer. *Nat. Plants* **4**, 116–127 (2018).
17. Zhang, H., Whitelegge, J. P. & Cramer, W. A. Ferredoxin:NADP$^+$ oxidoreductase is a subunit of the chloroplast cytochrome b_6f complex. *J. Biol. Chem.* **276**, 38159–38165 (2001).
18. Zhu, X. G., Long, S. P. & Ort, D. R. Improving photosynthetic efficiency for greater yield. *Annu. Rev. Plant Biol.* **61**, 235–261 (2010).
19. Simkin, A. J., McAusland, L., Lawson, T. & Raines, C. A. Overexpression of the RieskeFeS protein increases electron transport rates and biomass yield. *Plant Physiol.* **175**, 134–145 (2017).
20. Zhang, Z. et al. Electron transfer by domain movement in cytochrome bc_1. *Nature* **392**, 677–684 (1998).
21. Yan, J., Kurisu, G. & Cramer, W. A. Intraprotein transfer of the quinone analogue inhibitor 2,5-dibromo-3-methyl-6-isopropyl-*p*-benzoquinone in the cytochrome b_6f complex. *Proc. Natl Acad. Sci. USA* **103**, 69–74 (2006).
22. Yamashita, E., Zhang, H. & Cramer, W. A. Structure of the cytochrome b_6f complex: quinone analogue inhibitors as ligands of heme c_n. *J. Mol. Biol.* **370**, 39–52 (2007).
23. Hasan, S. S., Yamashita, E., Baniulis, D. & Cramer, W. A. Quinone-dependent proton transfer pathways in the photosynthetic cytochrome b_6f complex. *Proc. Natl Acad. Sci. USA* **110**, 4297–4302 (2013).
24. Sarewicz, M. et al. Metastable radical state, nonreactive with oxygen, is inherent to catalysis by respiratory and photosynthetic cytochromes bc_1/b_6f. *Proc. Natl Acad. Sci. USA* **114**, 1323–1328 (2017).
25. Singh, S. K. et al. Trans-membrane signalling in photosynthetic state transitions: redox- and structure-dependent interaction in vitro between Stt7 kinase and the cytochrome b_6f complex. *J. Biol. Chem.* **291**, 21740–21750 (2016).
26. Zito, F., Vinh, J., Popot, J. L. & Finazzi, G. Chimeric fusions of subunit IV and PetL in the b_6f complex of *Chlamydomonas reinhardtii*: structural implications and consequences on state transitions. *J. Biol. Chem.* **277**, 12446–12455 (2002).
27. Saif Hasan, S., Yamashita, E. & Cramer, W. A. Transmembrane signaling and assembly of the cytochrome b_6f-lipidic charge transfer complex. *Biochim. Biophys. Acta* **1827**, 1295–1308 (2013).
28. Nawrocki, W. J. et al. The mechanism of cyclic electron flow. *Biochim. Biophys. Acta* **1860**, 433–438 (2019).
29. Osyczka, A., Moser, C. C., Daldal, F. & Dutton, P. L. Reversible redox energy coupling in electron transfer chains. *Nature* **427**, 607–612 (2004).
30. Świerczek, M. et al. An electronic bus bar lies in the core of cytochrome bc_1. *Science* **329**, 451–454 (2010).
31. Alric, J., Pierre, Y., Picot, D., Lavergne, J. & Rappaport, F. Spectral and redox characterization of the heme c_i of the cytochrome b_6f complex. *Proc. Natl Acad. Sci. USA* **102**, 15860–15865 (2005).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Complex purification

Dimeric cytb_6f was isolated from dark-adapted market spinach (*S. oleracea*) in a procedure adapted from Dietrich and Kuhlbrandt[32].

In brief, spinach leaves were homogenized in buffer 1 (50 mM Tris-HCl pH 7.5, 200 mM sucrose and 100 mM NaCl). Homogenate was then filtered and centrifuged for 15 min at 4,540*g*, 4 °C. Following centrifugation, the supernatant containing cell debris was discarded and the pellet resuspended in buffer 2 (10 mM Tricine-NaOH pH 8 and 150 mM NaCl) before centrifugation again for 15 min (4,540*g*, 4 °C). The resultant pellet was resuspended in buffer 3 (2 M NaBr, 10 mM Tricine-NaOH pH 8 and 300 mM sucrose) and incubated on ice for 15 min before diluting twofold with ice-cold milliQ H_2O and centrifuging (15 min, 4,540*g*, 4 °C). The resultant pellet was resuspended in buffer 3 and incubated on ice for 15 min before diluting twofold with ice-cold milliQ H_2O and centrifuging again (15 min, 4,540*g*, 4 °C). The pellet was resuspended in buffer 2 and centrifuged for 15 min, 4,540*g* at 4 °C. The final pellet was resuspended in a small volume of buffer 4 (40 mM Tricine pH 8.0, 10 mM $MgCl_2$ and 10 mM KCl). The resultant thylakoid suspension was adjusted to 10 mg ml^{-1} chlorophyll (chlorophyll concentrations determined as described previously[33]).

For selective solubilization of cytb_6f, the thylakoid suspension (10 mg ml^{-1} chlorophyll) was diluted with membrane extraction buffer (40 mM Tricine pH 8.0, 10 mM $MgCl_2$, 10 mM KCl and 1.25% (w/v) Hecameg) to a final concentration of 2 mg ml^{-1} chlorophyll, 1% (w/v) Hecameg. The resultant solution was mixed thoroughly then incubated for 2 min at room temperature before dilution to 0.75% (w/v) Hecameg with buffer 4. Unsolubilized material was removed by ultracentrifugation at 244,000*g* at 4 °C for 30 min in a Beckman Ti50.2 rotor.

The solubilization supernatant was concentrated using a Centriprep 100K centrifugal filter (Merck Millipore) before loading onto a 10–40% (w/v) continuous sucrose gradient containing 40 mM Tricine pH 8, 10 mM $MgCl_2$, 10 mM KCl, 0.8% (w/v) Hecameg and 0.1 mg ml^{-1} egg yolk L-α-phosphatidylcholine (Sigma). This was ultracentrifuged at 174,587*g* at 4 °C for 16 h in a Beckman SW32 rotor.

A brown-ish band containing cytb_6f was collected from a region of the gradient corresponding to ~16% sucrose. This band was concentrated and loaded onto a ceramic hydroxyapatite column (CHT) (Type I, Bio-Rad) pre-equilibrated in 20 mM Hecameg, 0.1 mg ml^{-1} phosphatidylcholine and 20 mM Tricine pH 8. The column was washed with 5 column volumes of CHT wash buffer (20 mM Hecameg, 0.1 mg ml^{-1} phosphatidylcholine and 100 mM ammonium phosphate pH 8) before bound material was eluted with CHT elution buffer (20 mM Hecameg, 0.1 mg ml^{-1} phosphatidylcholine and 400 mM ammonium phosphate pH 8).

Detergent exchange and gel filtration

Concentrated CHT eluate was loaded onto a 10–35% (w/v) continuous sucrose gradient containing 50 mM HEPES pH 8, 20 mM NaCl and 0.3 mM 4-*trans*-(4-*trans*-propylcyclohexyl)-cyclohexyl α-maltoside (tPCCαM) and ultracentrifuged at 175,117*g* at 4 °C for 16 h in a Beckman SW41 rotor.

A single brown band containing cytb_6f was collected from a region of the gradient corresponding to ~22% sucrose. This band was concentrated and loaded onto HiLoad 16/600 Superdex 200 pg gel filtration column (GE Healthcare) connected to an ÄKTA prime plus purification system (GE Healthcare). The column was run at a rate of 0.2 ml min^{-1} with 145 ml with gel filtration buffer (50 mM HEPES pH 8, 20 mM NaCl, 0.3 mM tPCCαM). Eluted fractions comprising dimeric cytb_6f were pooled and concentrated.

SDS–PAGE and BN-PAGE analysis

Samples collected from each purification step were analysed by SDS–PAGE and BN-PAGE. For SDS–PAGE, precast NuPAGE 12% Bis-Tris gels (Invitrogen) were run for 60 min at 180 V before staining with Coomassie blue. For BN-PAGE, precast NativePAGE 3–12% Bis-Tris gels (Invitrogen) were run for 120 min at 160 V before staining with Coomassie blue. Gels were imaged using an Amersham 600 imager (GE Healthcare).

Quantification of purified dimeric cytb_6f using redox difference spectra

Absorbance spectra were recorded at room temperature on a Cary60 spectrophotometer (Agilent). For redox difference spectra cytochromes were first fully oxidized with a few grains of potassium ferricyanide followed by reduction with a few grains of sodium ascorbate (cyt*f*) then sodium dithionite (cyt*f* and cytb_6). At each stage the sample was mixed thoroughly and incubated for ~1 min before recording spectra. Redox difference spectra (ascorbate-reduced minus ferricyanide-oxidized and dithionite-reduced minus ascorbate-reduced) were calculated and used to determine the concentrations of *c*-type haem of cyt*f* and the two *b*-type haems of cytb_6 using extinction coefficients of 25 mM cm^{-1} (haem *f*) and 21 mM cm^{-1} (cytb_6 haems)[34].

Reduction of decylplastoquinone

Approximately 0.1 mg decylplastoquinone (Merck) was dissolved in 100 µl ethanol, mixed with a few grains of sodium dithionite dissolved in 100 µl milliQ H_2O and vortexed until the solution became colourless. Decylplastoquinol was extracted by mixing with 0.5 ml hexane, vortexing and centrifuging at 16,000*g* for 2 min. The hexane layer was carefully removed ensuring none of the aqueous phase was collected. Hexane extraction was repeated on the aqueous phase twice more, then the hexane solutions were pooled and dried in a rotary evaporator at 30 °C for 1 h before re-dissolving in ~100 µl DMSO. Decylplastoquinol concentration was determined by diluting 10 µl of the DMSO solution into 795 µl ethanol, recording the absorbance spectrum between 250 and 350 nm and using an extinction coefficient[35] of 3,540 M^{-1} cm^{-1} at 290 nm.

Purification of PC

PC was purified in its oxidized form from market spinach. In brief, spinach leaves were homogenized in buffer containing 50 mM sodium phosphate pH 7.4, 5 mM $MgCl_2$ and 300 mM sucrose. Homogenate was then filtered and centrifuged for 15 min at 4,000*g*. Following centrifugation, the supernatant containing cell debris was discarded and the pellet was resuspended in buffer containing 10 mM Tricine pH 7.4 and 5 mM $MgCl_2$. The solution was incubated on ice for 1 min before diluting twofold with buffer containing 10 mM Tricine pH 7.4, 5 mM $MgCl_2$, 400 mM sucrose and centrifuging for 15 min at 4,000*g*. Following centrifugation, the pellet was resuspended to a chlorophyll concentration of 2 mg ml^{-1} in buffer containing 10 mM HEPES pH 7.6, 5 mM NaCl and 5 mM EDTA, and sonicated for 10 min, at 30 s intervals. The solution was centrifuged at 200,000*g* for 1 h to pellet any large unbroken material. The supernatant was applied to four 5-ml GE Healthcare Hi-TRAP Q FF anion-exchange columns in series, equilibrated in HEPES pH 8, 5 mM NaCl. A gradient of 0.005–1 M NaCl was used for elution, with PC eluting at around 200 mM. PC-containing fractions were identified by the blue colour on addition of potassium ferricyanide. These fractions were pooled, concentrated in a Vivaspin 3-kDa molecular-weight cut-off spin concentrator and loaded onto a Superdex 200 16/600 FPLC column, equilibrated with 20 mM HEPES pH 8 and 20 mM NaCl. The resulting PC fractions were pooled, concentrated and frozen at −80 °C until use.

Activity assays

Reduction of PC by cytb_6f was monitored by stopped-flow absorbance spectroscopy using an Olis RSM 1000 rapid-scanning

spectrophotometer equipped with a USA-SF stopped flow cell at 20 °C. Solution A (231.25 nM cytb_6f and 62.5 μM PC in 50 mM HEPES pH 8, 20 mM NaCl and 0.3 mM tPCCαM) and solution B (1.25 mM decylplastoquinol in the same buffer) were prepared and the reaction was initiated by mixing the solutions in a 4:1 volumetric ratio (final concentrations: 185 nM cytb_6f, 50 μM PC and 250 μM decylplastoquinol). PC reduction was monitored by recording absorbance spectra between 420 and 750 nm at a rate of 62 scans s^{-1} and plotting the change in absorbance[36] at 597 nm. In a control reaction, cytb_6f was omitted to record the uncatalysed reduction of PC by decylplastoquinol. Fitting of the initial reaction rates was performed in Origin. All measurements were carried out in triplicate.

CryoEM specimen preparation and data acquisition

In brief, 3 μl of purified cytb_6f (~17 μM) was applied to freshly glow-discharged holey carbon grids (Quantifoil R1.2/1.3, 400 mesh Cu). The grids were blotted for 2 s at 8 °C then plunge frozen into liquid ethane using a Leica EM GP at 90% relative humidity. Data acquisition was carried out on a Titan Krios microscope operated at 300 kV (Thermo Fisher) equipped with an energy filtered (slit width 20 eV) K2 summit direct electron detector. A total of 6,035 movies were collected in counting mode at a nominal magnification of 130,000× (pixel size of 1.065 Å) and a dose of 4.6 e^- $Å^{-2}$ s^{-1} (see Extended Data Table 1). An exposure time of 12 s was used and the resulting movies were dose-fractionated into 48 fractions. A defocus range of −1.5 to −2.5 μm was used.

Image processing and 3D reconstruction

Beam-induced motion correction and dose-fractionation were carried out using MotionCor2. Contrast transfer function (CTF) parameters of the dose-weighted motion-corrected images were then estimated using GCTF[37]. All subsequent processing steps were performed using RELION 2.1[38] or 3.0[39] unless otherwise stated.

In total, 422,660 particles were manually picked from 6,035 micrographs. These particles were extracted using a box size of 220 × 220 pixels and subjected to reference-free 2D classification. A typical micrograph showing picked particles is shown in Extended Data Fig. 2a, b. Particles that categorized into poorly defined classes were rejected, while the remaining 292,242 (69.2%) particles were used for further processing. A subset of 30,000 particles was used to generate a de novo initial model using the '3D initial model' subroutine. The initial model low-pass filtered to 20 Å was used as a reference map for subsequent 3D classification into 10 3D classes. One stable 3D class at a resolution of 5.38 Å was selected for high-resolution 3D auto-refinement; this class accounted for a subset of 108,560 particles (25.6%). This subset of refined particles was then re-extracted and re-centred before another round of 3D auto-refinement was carried out. The resultant 4.85 Å density map was corrected for the modulation transfer function (MTF) of the Gatan K2 summit camera then further sharpened using the post-processing procedure to 4.02 Å. Per-particle CTF-refinement was carried out and a soft mask was created which included the detergent shell. The final global resolution estimate of 3.58 Å was based on the gold-standard Fourier shell correlation (FSC) cut-off of 0.143.

Local resolution was determined using one of two unfiltered half-maps as an input, a calibrated pixel size of 1.065 and a B-factor of −103. The output local resolution map is shown in Extended Data Fig. 2d, e.

Model building

Initially, a homology-based approach was performed using the crystallographic structure of *Nostoc* sp. PCC 7120 cytb_6f (PDB: 4OGQ)[40] as a template. Sequence alignments of the eight polypeptide subunits of cytb_6f were carried out using Clustal Omega (Extended Data Figs. 7, 8). The model was rigid-body docked into the density using the 'fit in map' tool in Chimera[41]. This was then followed by manual adjustment and real-space refinement using COOT[42]. Sequence assignment and fitting was guided by bulky residues such as Arg, Trp, Tyr and Phe. After fitting of the polypeptide chains and cofactors in one half of the dimeric complex, the other half of the complex was then independently fitted into the C1 density map. Once both halves of the complex were fitted, cofactors, lipids and plastoquinone-9 molecules were fitted into regions of unassigned density. The final model underwent global refinement and minimization using the real space refinement tool in PHENIX[43]. The final refinement statistics are summarized in Extended Data Table 1.

Pigment analysis by reversed-phase HPLC

Pigments were extracted from purified cytb_6f with 7:2 acetone:methanol (v/v) and clarified extracts were separated by reversed-phase HPLC at a flow rate of 1 ml min^{-1} at 40 °C using a Supelco Discovery HS C18 column (5 μm particle size, 120 Å pore size, 25 cm × 4.6 mm) on an Agilent 1200 HPLC system. The column was equilibrated in acetonitrile: water:trimethylamine (9:1:0.01 v/v/v) and pigments were eluted by applying a linear gradient of 0–100% ethyl acetate over 15 min followed by isocratic elution with 100% ethyl acetate for a further 5 min. Elution of carotenoid and chlorophyll species was monitored by absorbance at 400, 450, 490 and 665 nm. Chlorophyll *a* was identified by its absorption spectra and known retention time[44]. The major carotenoid species was confirmed as 9-*cis* β-carotene using a standard obtained from Sigma-Aldrich (product no. 52824).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All relevant data are available from the authors and/or are included with the manuscript or in the Supplementary Information. Atomic coordinates and the cryo-EM density map have been deposited in the Protein Data Bank under accession number 6RQF and the Electron Microscopy Data Bank (EMDB) under accession number EMD-4981.

32. Dietrich, J. & Kühlbrandt, W. Purification and two-dimensional crystallization of highly active cytochrome b_6f complex from spinach. *FEBS Lett.* **463**, 97–102 (1999).
33. Porra, R. J., Thompson, W. A. & Kriedemann, P. E. Determination of accurate extinction coefficients and simultaneous equations for assaying chlorophylls *a* and *b* extracted with four different solvents: verification of the concentration of chlorophyll standards by atomic absorption spectroscopy. *Biochim. Biophys. Acta* **975**, 384–394 (1989).
34. Cramer, W. A. & Whitmarsh, J. Photosynthetic cytochromes. *Annu. Rev. Plant Physiol.* **28**, 133–172 (1977).
35. Dawson, R. M. C., Elliot, D. C., Elliot, W. H. & Jones, K. M. *Data for Biochemical Research* (Clarendon, 1986).
36. Tan, S. & Ho, K. K. Purification of an acidic plastocyanin from *Microcystis aeruginosa*. *Biochim. Biophys. Acta* **973**, 111–117 (1989).
37. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
38. Fernandez-Leiro, R. & Scheres, S. H. W. A pipeline approach to single-particle processing in RELION. *Acta Crystallogr. D* **73**, 496–502 (2017).
39. Zivanov, J. et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7**, e42166 (2018).
40. Hasan, S. S. & Cramer, W. A. Internal lipid architecture of the hetero-oligomeric cytochrome b_6f complex. *Structure* **22**, 1008–1015 (2014).
41. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
42. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
43. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
44. Proctor, M. S. et al. Plant and algal chlorophyll synthases function in *Synechocystis* and interact with the YidC/Alb3 membrane insertase. *FEBS Lett.* **592**, 3062–3073 (2018).
45. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand and contrast loss in single particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).

Acknowledgements M.P.J. acknowledges funding from the Leverhulme Trust grant RPG-2016-161. C.N.H., P.Q., A.H., D.J.K.S. and M.P.J. also acknowledge financial support from the Biotechnology and Biological Sciences Research Council (BBSRC UK) award numbers BB/M000265/1 and BB/P002005/1. L.A.M. was supported by a White Rose doctoral studentship, G.E.M. was supported by a doctoral studentship from The Grantham Foundation and D.A.F. was supported by a University of Sheffield doctoral scholarship. Cryo-EM data was collected at

the Astbury Biostructure Laboratory funded by the University of Leeds (ABSL award) and the Wellcome Trust (108466/Z/15/Z). We thank S. Tzokov, J. Bergeron, J. Wilson and D. Mann for their helpful advice and assistance with the EM and data processing.

Author contributions P.Q., C.N.H., N.A.R. and M.P.J. supervised the project. L.A.M., G.E.M., P.Q., C.N.H., R.F.T. and M.P.J. designed the experiments. L.A.M. and G.E.M. purified the cytb_6f complex, L.A.M., G.E.M., A.H. and D.J.K.S. characterized the cytb_6f complex. L.A.M., P.Q., D.A.F. and R.F.T. collected, processed and/or analysed the cryo-EM data. L.A.M., C.N.H. and M.P.J. wrote the manuscript. All authors proofread and approved the manuscript.

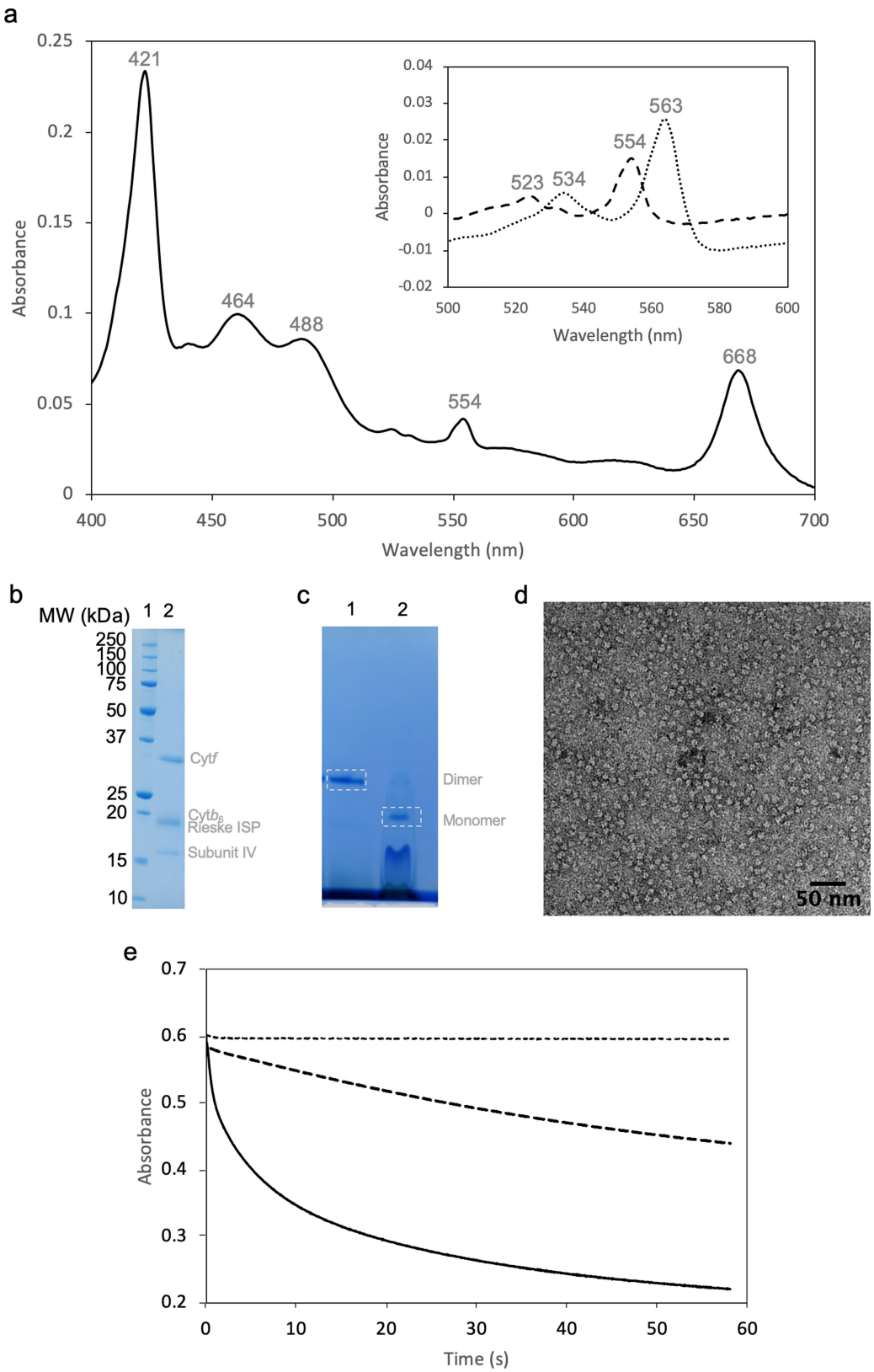
Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1746-6.

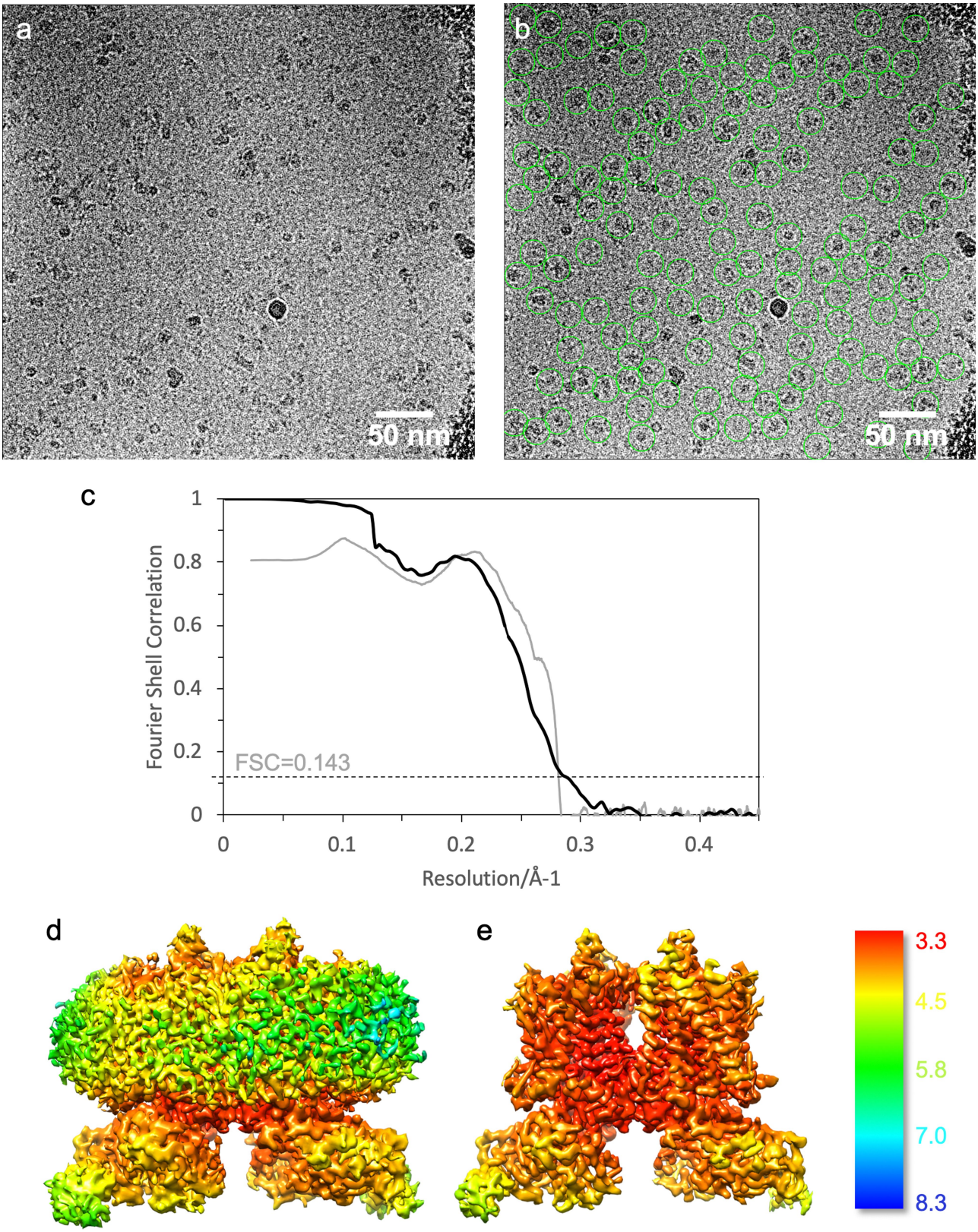
Correspondence and requests for materials should be addressed to C.N.H. or M.P.J.

Peer review information *Nature* thanks Zhenfeng Liu, Alexander Tikhonov and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

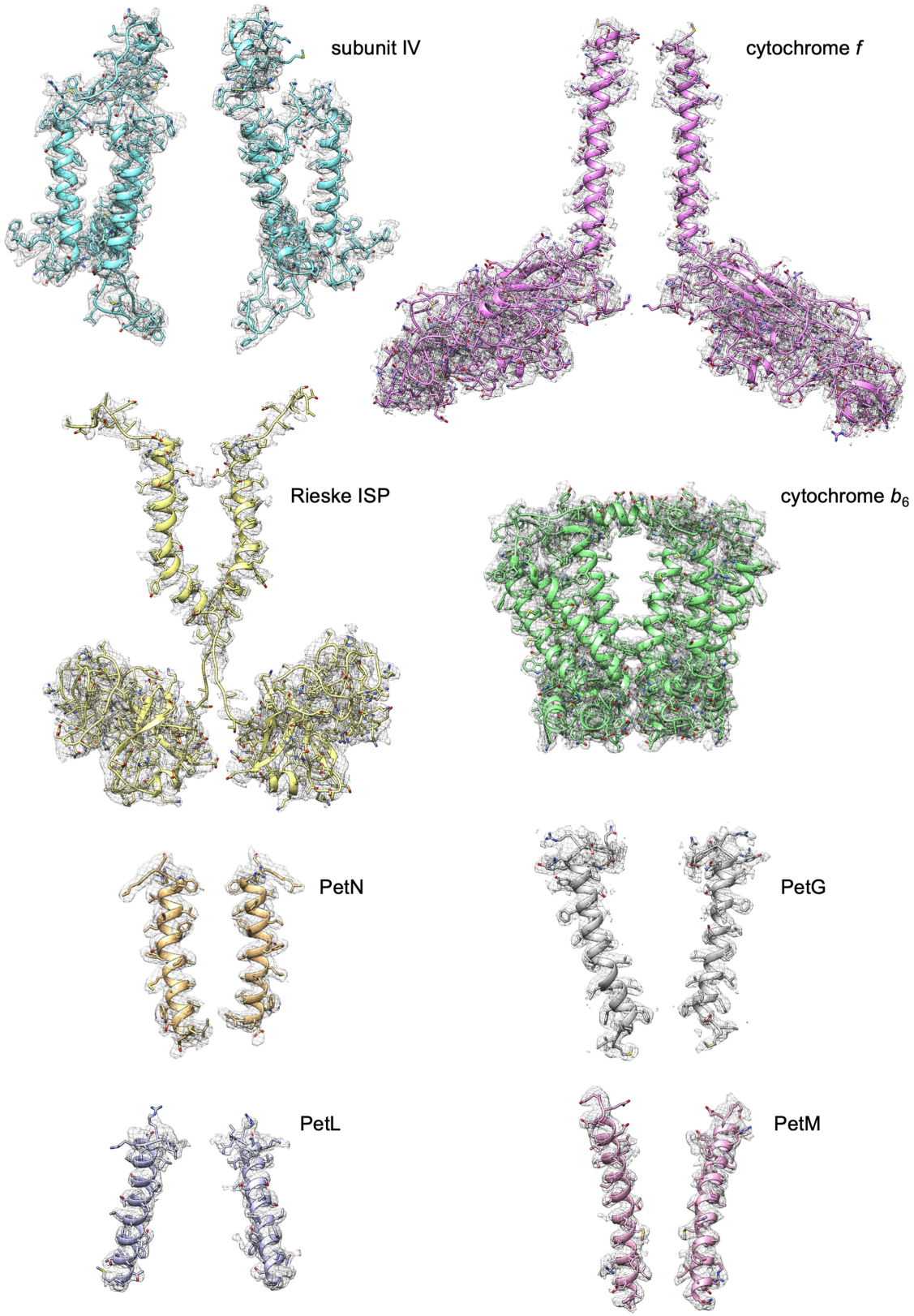


Extended Data Fig. 1 | See next page for caption.

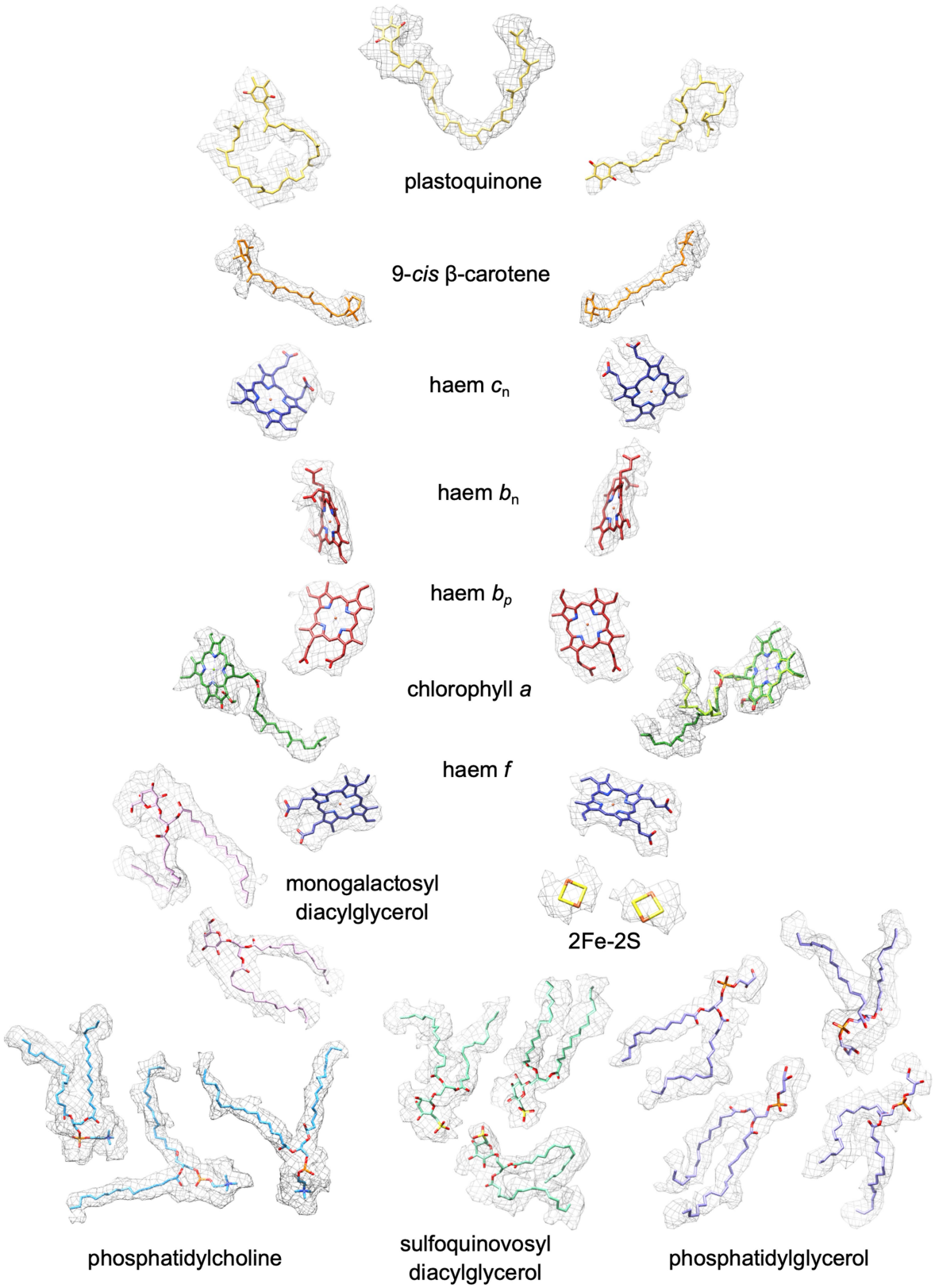
Extended Data Fig. 1 | Purification of cytb_6f from spinach. a, Absorption spectrum of ascorbate-reduced purified b_6f complex. The peak at 421 nm corresponds to the Soret band of bound pigments (chlorophyll *a* and haems). The peaks at 554 and 668 nm correspond to *c*-type haems of cyt*f* and chlorophyll *a*, respectively. The inset panel shows redox difference spectra of ascorbate-reduced minus ferricyanide-oxidized b_6f (dashed line) and dithionite-reduced minus ascorbate-reduced (dotted line) cytb_6f. Redox difference spectra show haem *f* absorption peaks at 523 and 554 nm as well as absorption peaks at 534 and 563 nm corresponding to the *b*-type haems of cytb_6. The calculated ratio of cytb_6 *b*-type haems to the *c*-type haem of cyt*f* was ~2 using extinction coefficients of 25 mM cm^{-1} (*f*) and 21 mM cm^{-1} (b_6)[34]. The spectra exhibit the absorption properties characteristic of intact cytb_6f. Spectra were recorded at room temperature. **b**, SDS–PAGE analysis of purified cytb_6f indicates that the sample is highly pure, with the four large subunits of the complex (cyt*f*, cytb_6, the Rieske ISP and subunit IV) running at ~31 kDa, ~24 kDa, ~20 kDa and ~17 kDa, respectively and the four small subunits (PetG, PetL, PetM and PetN) running at around 4 kDa (not shown). **c**, **d**, Negative-stain and BN-PAGE analysis of purified cytb_6f demonstrates the sample is dimeric and highly homogenous, with a single band corresponding to dimeric cytb_6f shown in lane 1. Lane 2 shows a sample that has been deliberately monomerized following incubation with 1% Triton X-100 for 1 h. For gel source data see Supplementary Fig. 1. **e**, The catalytic rate of plastocyanin reduction by the purified dimeric cytb_6f complex as determined by stopped-flow absorbance spectroscopy. A rate of 200 e^- s^{-1} was determined by taking the initial linear region from the enzyme-catalysed reaction (solid line) and subtracting the background rate measured in the absence of enzyme (long-dashed line). Plastocyanin reduction was not observed in the absence of decylplastoquinol (short-dashed line). Reactions were initiated upon addition of decylplastoquinol to the solution containing plastocyanin and b_6f while monitoring the loss of absorbance at 597 nm. Final concentrations were 50 μM plastocyanin, 185 nm b_6f and 250 μM decylplastoquinol. All experiments were performed in triplicate and controls were performed in the absence of b_6f or decylplastoquinol.



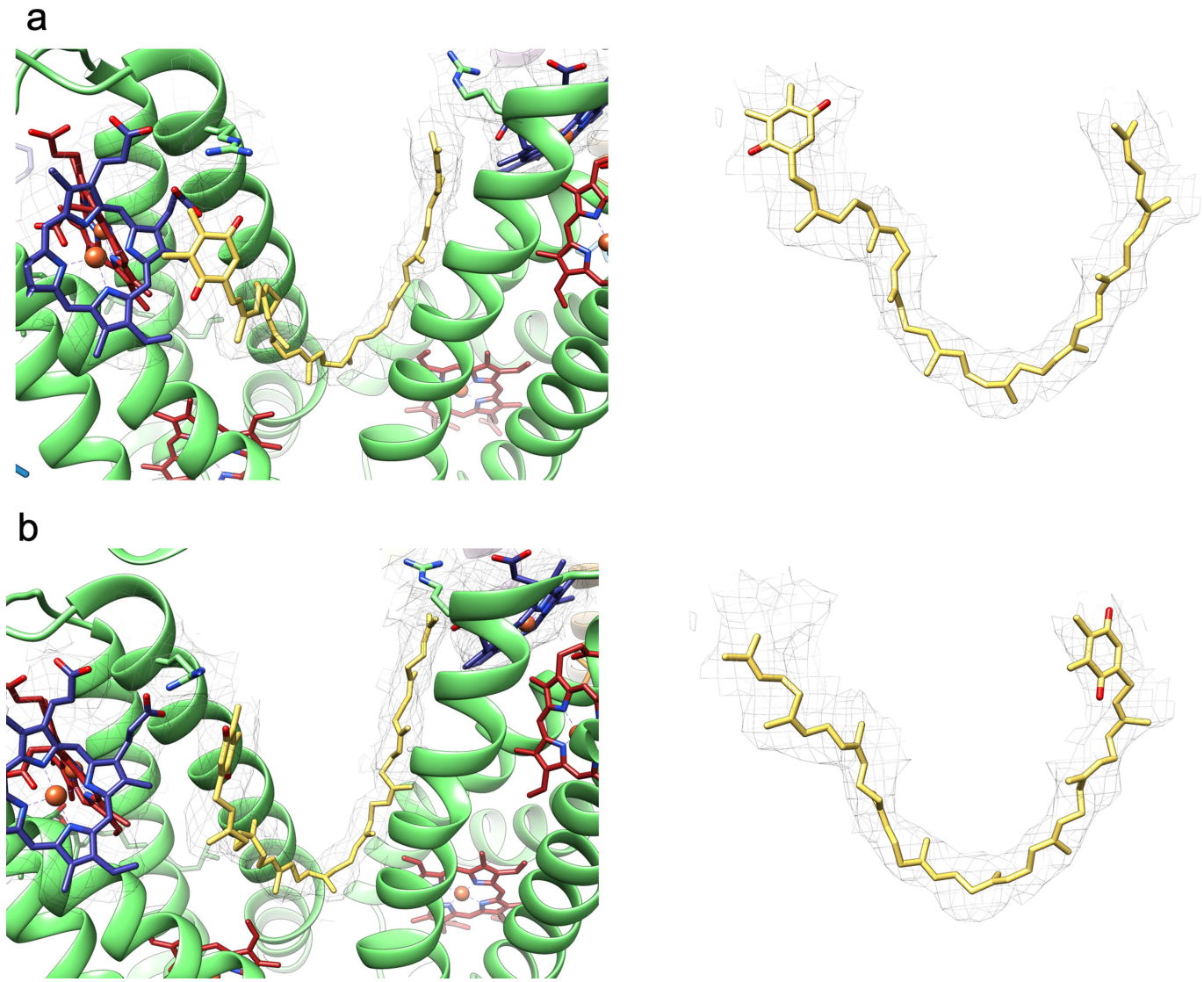
Extended Data Fig. 2 | Cryo-EM micrographs of the spinach cytb_6f complex and calculation of the cryo-EM map global and local resolution. a, Cytb_6f particles covered by a thin layer of vitreous ice on a supported carbon film. **b**, Examples of dimeric cytb_6f particles are circled in green. We recorded 6,035 cryo-EM movies, from which 422,660 particles were picked manually for reference-free 2D classification. The final density map was calculated from 108,560 particles. **c**, Gold-standard refinement was used for estimation of the final map resolution (solid black line). The global resolution of 3.58 Å was calculated using a FSC cut-off at 0.143. A model-to-map FSC curve (solid grey line) was also calculated. **d**, **e**, A C1 density map of the cytb_6f complex both with (**d**) and without (**e**) the detergent shell. The map is coloured according to local resolution estimated by RELION and viewed from within the plane of the membrane. The colour key on the right shows the local structural resolution in angstroms (Å).



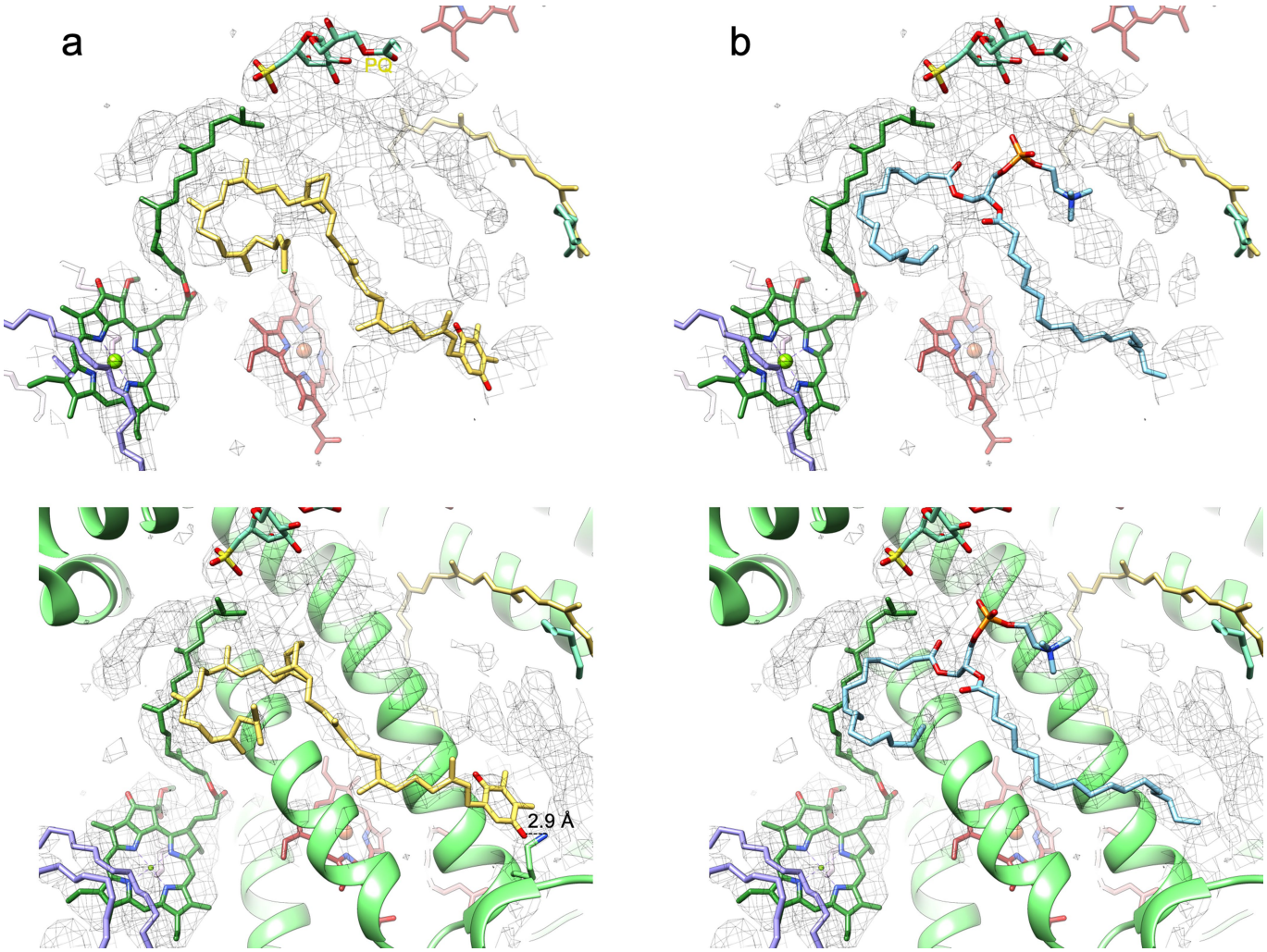
Extended Data Fig. 3 | Cryo-EM densities and structural models of polypeptides in the cytb_6f complex. Polypeptides are coloured as in Fig. 1. The contour levels of the density maps were adjusted to 0.0144.



Extended Data Fig. 4 | Cryo-EM densities and structural models of prosthetic groups, lipids and plastoquinone molecules in the cytb_6f complex. *c*-type haems (*f*, c_n; dark blue), *b*-type haems (b_p, b_n; red), 9-*cis* β-carotene (orange), chlorophyll *a* (major conformation, dark green; minor conformation, light green), 2Fe-2S (burnt orange and yellow), plastoquinones (yellow), monogalactosyl diacylglycerol (light pink), phosphatidylcholine (light cyan), sulfoquinovosyl diacylglycerol (light green) and phosphatidylglycerol (light purple). The contour levels of the density maps were adjusted to 0.0068.



Extended Data Fig. 5 | Alternative interpretation of the region assigned as PQ2. a, **b**, The density map showing two possible alternative conformations for PQ2, the major conformation (**a**) and the alternative conformation (**b**). Cofactors are coloured as in Extended Data Fig. 4 with *b*-type haems (b_p and b_n) coloured red, *c*-type haems (c_n) coloured dark blue, chlorophyll *a* (major conformation) coloured dark green, plastoquinones coloured yellow and the cytb_6 subunit coloured light green. The contour level of the density map was adjusted to 0.0089.



Extended Data Fig. 6 | Alternative interpretations of the density map in the region assigned as PQ3. a, **b**, The density map modelled with a plastoquinone molecule (**a**) and a phosphatidylcholine molecule (**b**). Top, the protein-free density map; bottom, the map including cytb_6 (green). The 2.9 Å distance indicates a close contact between the PQ3 head group and the conserved Lys208. Cofactors are coloured as in Extended Data Fig. 4 with *b*-type haems (b_p and b_n) coloured red, chlorophyll *a* (major conformation) in dark green, plastoquinones in yellow, phosphatidylcholine in light cyan, sulfoquinovosyl diacylglycerol in mint green and the cytb_6 subunit in light green. The contour level of the density map was adjusted to 0.0127.

a

```
Mastigocladus     YPFWAQQTYPETPREPTGRIVCANCHLAAKPAEVEVPQSVLPDTVFKAVVKIPYDTKLQQ 60
Nostoc            YPFWAQQTYPETPREPTGRIVCANCHLAAKPTEVEVPQSVLPDTVFKAVVKIPYDTSVQQ 60
Chlamydomonas     YPVFAQQNYA-NPREANGRIVCANCHLAQKAVEIEVPQAVLPDTVFEAVIELPYDKQVKQ 59
Spinach           YPIFAQQGYE-NPREATGRIVCANCHLANKPVDIEVPQAVLPDTVFEAVVRIPYDMQLKQ 59
                  **.:*** *  .*** .********** * .::*****:*******:**:.:*** .::*

Mastigocladus     VAADGSKVGLNVGAVLMLPEGFKIAPEERIPEELKKEVGDVYFQPYKEGQDNVLLVGPLP 120
Nostoc            VGADGSKVGLNVGAVLMLPEGFKIAPEDRIPEELKEEIGDVYFQPYGEDKDNIVIVGPLP 120
Chlamydomonas     VLANGKKGDLNVGMVLILPEGFELAPPDRVPAEIKEKVGNLYYQPYSPEQKNILVVGPVP 119
Spinach           VLANGKKGGLNVGAVLILPEGFELAPPDRISPEMKEKMGNLSFQSYRPNKQNILVIGPVP 119
                  * *:*.* .**** **:*****::** :*:  *:*::::*:: :* *   :.*::::**:*

Mastigocladus     GEQYQEIVFPVLSPNPTTDKNIHFGKYAIHLGANRGRGQIYPTGEKSNNNVFTASATGTI 180
Nostoc            GEQYQEIVFPVLSPNPANDKNIHFGKYSVHVGGNRGRGQVYPTGEKSNNNLYSAAATGTI 180
Chlamydomonas     GKKYSEMVVPILSPDPAKNKNVSYLKYPIYFGGNRGRGQVYPDGKKSNNTIYNASAAGKI 179
Spinach           GQKYSEITFPILAPDPATKKDVHFLKYPIYVGGNRGRGQIYPDGSKSNNTVYNSTATGIV 179
                  *:::*.*:..*:*:*:*:..*:: : ** ::.*.******:** *.****.::.::*:* :

Mastigocladus     TKIAKEEDEYGNVKYQVSIQTDSGKTVVDTIPAGPELIVSEGQAVKAGEALTNNPNVGGF 240
Nostoc            SKIAKQEGEDGSVKYLVDIKTESGEVVSDTIPAGPELIVSEGQAVTAGDALTNNPNVGGF 240
Chlamydomonas     VAITALSEKKGGFEVSIE-K-ANGEVVVDKIPAGPDLIVKEGQTVQADQPLTNNPNVGGF 237
Spinach           KKI--VRKEKGGYEINIA-DASDGREVVDIIPRGPELLVSEGESIKLDQPLTSNPNVGGF 236
                    *     : *. :  :  .  .*. * * ** **:*:*.**:::  .: **.*******

Mastigocladus     GQDDTEIVLQDPNRVKWMIAFICLVMLAQLMLILKKKQVEKVQAAEMNF    289
Nostoc            GQLDAEIVLQDANRVGWLIAFVALVMLAQVMLVLKKKQVEKVQAAEMNF    289
Chlamydomonas     GQAETEIVLQNPARIQGLLVFFSFVLLTQVLLVLKKKQFEKVQLAEMNF    286
Spinach           GQGDAEVVLQDPLRIQGLLFFFASVILAQIFLVLKKKQFEKVQLSEMNF    285
                  ** ::*:***:  *:  :: *.. *:*:*::*:*****.**** :****
```

b

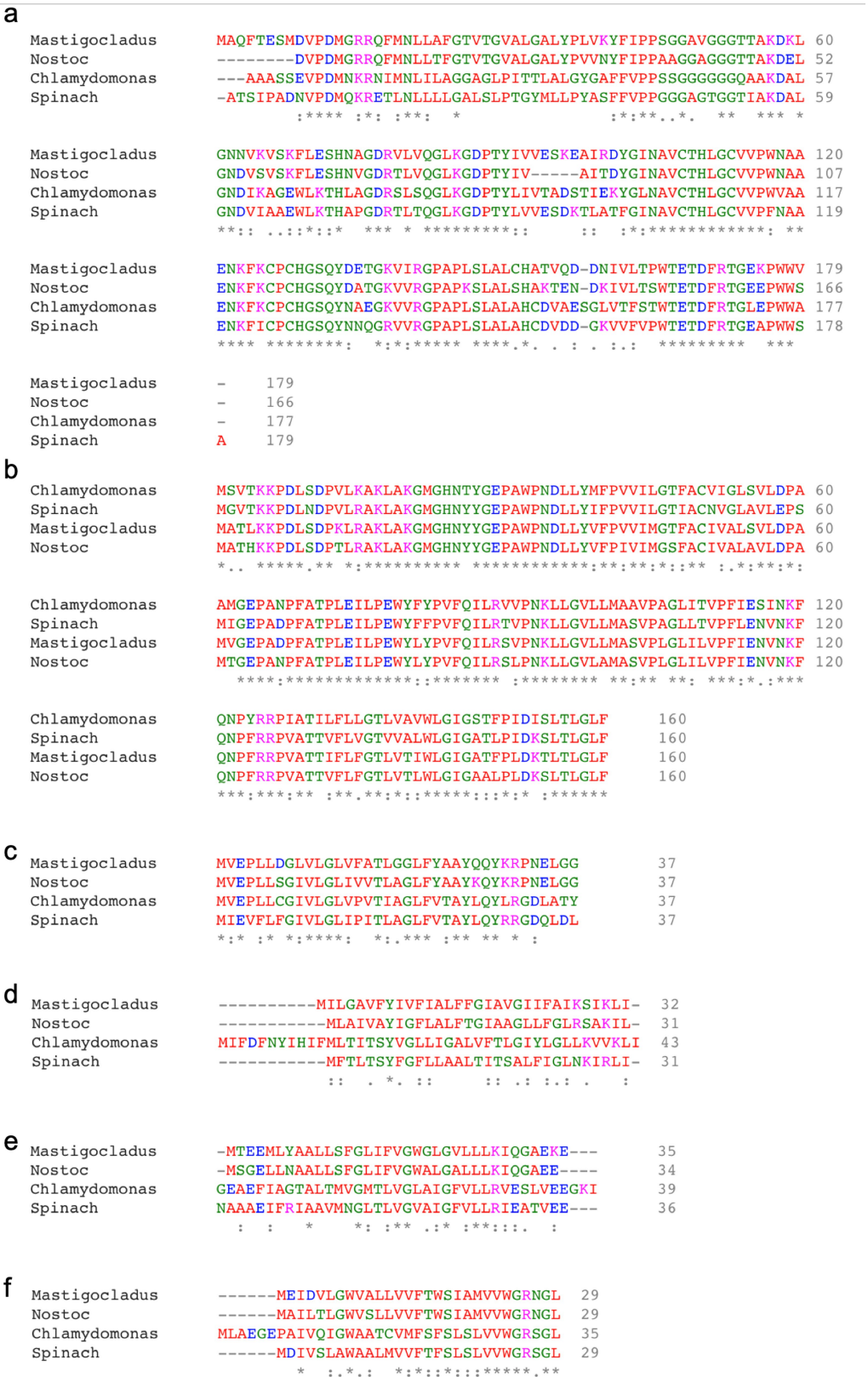
```
Mastigocladus     MANVYDWFQERLEIQALADDVTSKYVPPHVNIFYCLGGITLTCFLIQFATGFAMTFYYKP 60
Nostoc            MANVYDWFEERLEIQAIAEDVTSKYVPPHVNIFYCLGGITLVCFLIQFATGFAMTFYYKP 60
Chlamydomonas     MSKVYDWFEERLEIQAIADDITSKYVPPHVNIFYCIGGITFTCFLVQVATGFAMTFYYRP 60
Spinach           MSKVYDWFEERLEIQAIADDITSKYVPPHVNIFYCLGGITLTCFLVQVATGFAMTFYYRP 60
                  *::******:*******:*:*:***************:****:.***:*.**********:*

Mastigocladus     TVTEAYASVQYIMNEVSFGWLIRSIHRWSASMMVLMMILHVFRVYLTGGFKKPRELTWIS 120
Nostoc            TVAEAYSSVQYIMNEVNFGWLIRSIHRWSASMMVLMMILHVFRVYLTGGFKKPRELTWVS 120
Chlamydomonas     TVAEAFASVQYIMTDVNFGWLIRSIHRWSASMMVLMMVLHVFRVYLTGGFKRPRELTWVT 120
Spinach           TVTDAFASVQYIMTEVNFGWLIRSVHRWSASMMVLMMILHVFRVYLTGGFKKPRELTWVT 120
                  **:::*::******.:*.*******:***********:**********:******::

Mastigocladus     GVILAVITVSFGVTGYSLPWDQVGYWAVKIVSGVPEAIPVVGVLISDLLRGGSSVGQATL 180
Nostoc            GVILAVITVSFGVTGYSLPWDQVGYWAVKIVSGVPEAIPVVGVLISDLLRGGSSVGQATL 180
Chlamydomonas     GVIMAVCTVSFGVTGYSLPWDQVGYWAVKIVTGVPDAIPGVGGFIVELLRGGVGVGQATL 180
Spinach           GVVLGVLTASFGVTGYSLPWDQIGYWAVKIVTGVPDAIPVIGSPLVELLRGSASVGQSTL 180
                  **::.* *.*************:********:***:*** :*  : :****. .***:**

Mastigocladus     TRYYSAHTFVLPWLIAVFMLLHFLMIRKQGISGPL  215
Nostoc            TRYYSAHTFVLPWLIAVFMLFHFLMIRKQGISGPL  215
Chlamydomonas     TRFYSLHTFVLPLLTAVFMLMHFLMIRKQGISGPL  215
Spinach           TRFYSLHTFVLPLLTAVFMLMHFLMIRKQGISGPL  215
                  **:** ****** * *****:***************
```

Extended Data Fig. 7 | Multiple sequence alignment of cytb_6 *f* subunits cyt*f* and cytb_6. a, **b**, Sequences of cyt*f* (**a**) and cytb_6 (**b**) from cyanobacterial (*M. laminosus* and *Nostoc* sp. PCC7120), algal (*C. reinhardtii*) and plant (*S. oleracea*) subunits were aligned in Clustal Omega v.1.2.4. Conserved identities are indicated by asterisks, and similarities by double or single dots. Polar residues are coloured in green, positively charged residues are pink, hydrophobic residues are red and negatively charged residues are blue. The sequences omit signal peptides.



Extended Data Fig. 8 | Multiple sequence alignment of the Rieske ISP, subunit IV, PetG, PetL, PetM and PetN. a–**f**, Sequences of Rieske ISP (**a**), subunit IV (**b**), PetG (**c**), PetL (**d**), PetM (**e**) and PetN (**f**) from cyanobacterial (*M. laminosus* and *Nostoc* sp. PCC7120), algal (*C. reinhardtii*) and plant (*S. oleracea*) subunits were aligned in Clustal Omega v.1.2.4. Conserved identities are indicated by asterisks, and similarities by double or single dots. Polar residues are coloured in green, positively charged residues are pink, hydrophobic residues are red and negatively charged residues are blue. The sequences omit signal peptides.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

Cryo-EM data collection, refinement and validation statistics

	S. oleracea cytb_6f (EMD-4981) (PDB 6RQF)
Data collection and processing	
Magnification	130,000 X
Voltage (kV)	300
Electron exposure (e^{-}/Å^2)	1.15 (55.2 e- on 48 frames)
Defocus range (μm)	-1.5 to -2.5
Pixel size (Å)	1.065
Symmetry imposed	C1
Initial particle images (no.)	422,660
Final particle images (no.)	108,560
Map resolution (Å)	3.58
FSC threshold	0.143
Map resolution range (Å)	~3.3-8.3
Refinement	
Initial model used (PDB code)	RELION *de novo* model from 30,000 particles
Model resolution (Å)	3.58
FSC threshold	0.143
Model resolution range (Å)	~3.3-8.7
Map sharpening *B* factor (Å^2)	Estimated automatically using RELION*
Model composition	
Non-hydrogen atoms	16,359
Protein residues	1,944
Ligands	29
B factors (Å^2)	
Protein	RELION auto-estimated
Ligand	RELION auto-estimated
R.m.s. deviations (PHENIX)	
Bond lengths (Å)	0.009
Bond angles (°)	1.182
Validation	
MolProbity score	1.83
Clashscore	9.67
Poor rotamers (%)	0.18
Ramachandran plot	
Favored (%)	95.40
Allowed (%)	4.24
Disallowed (%)	0.37

*Data from ref. [45].

Extended Data Table 2 | Comparison of cofactor distances in b_6f and bc_1 dimers from different species

	(PDB 6RQF)	(PDB 1Q90)	(PDB 2E74)	(PDB 4OGQ)
Source	*S. oleracea*	*C. reinhardtii*	*M. laminosus*	*Nostoc sp. PCC 7120*
Resolution (Å)	3.6	3.1	3.0	2.5
Inhibitors *	-	TDS (Q_p)	-	-
Distances:				
b_n - c_n (Å)	4.7, 4.7	4.7, 4.7	4.7, 4.7	4.6, 4.6
b_n - b_p (Å)	12.1, 12.0	12.2, 12.2	12.2, 12.2	12.1, 12.1
b_p - b_p (Å)	15.3	15.1	15.2	15.3
b_p - [2Fe-2S] (Å)	25.6, 25.5	22.9, 22.9	25.5, 25.5	25.3, 25.3
[2Fe-2S] - f (Å)	25.9, 26.1	27.8, 27.8	26.2, 26.2	26.2, 26.2

	(PDB 1BCC) (distal)	(PDB 3BCC) (proximal)
Source	*G. gallus*	*G. gallus*
Resolution (Å)	3.2	3.7
Inhibitors *	-	STG (Q_p), AMY (Q_n)
Distances:		
b_n - b_p (Å)	12.4, 12.4	12.3, 12.3
b_p - b_p (Å)	14.4	14.5
b_p - [2Fe-2S] (Å)	30.3, 30.3	23.0, 23.1
[2Fe-2S] - c_1 (Å)	16.8, 16.8	27.3, 27.5

The distances are edge-to-edge (Å), for each half of the b_6f dimer from different species (PDB: 6RQF, 1Q90, 2E74, 4OGQ) and the bc_1 dimer from *Gallus gallus* with the Rieske ISP in its distal (PDB: 1BCC) and proximal (PDB: 3BCC) positions. AMY, antimycin; STG, stigmatellin.

nature research

Corresponding author(s): *Double-blind peer review submissions: write DBPR and your manuscript number here instead of author names.*

Last updated by author(s): *YYYY-MM-DD*

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Thermo Fisher EPU 1.9
Data analysis	Relion 2.1/3.0, MotionCorr2, GCTF, Coot, Pymol 2.3.1, Chimera 1.13.1, Phenix 1.15.2, CLUSTAL Omega 1.2.4, Origin Pro 2018, Microsoft Excel 16.27

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All relevant data are available from the authors and/or are included with the manuscript or Supplementary Information. Atomic coordinates and the cryo-EM density map have been deposited in the Protein Data Bank (PDB) under accession number 6RQF and the Electron Microscopy Data Bank (EMDB) under accession number EMD-4981.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. For cryoEM, sample sizes were determined by available electron microscopy time and density of particles on the electron microscopy grids. 422,660 particles were manually picked from 6,035 micrographs. Particles that categorised into poorly defined classes were rejected, while the remaining 292,242 (69.2%) particles were used for further processing. A subset of 30,000 particles was used to generate a de novo initial model using the '3D initial model' subroutine. The initial model low pass filtered to 20 Å was used as a reference map for subsequent 3D classification into 10 3D classes. One stable 3D class at a resolution of 5.38 Å was selected for high resolution 3D auto-refinement; this class accounted for a subset of 108,560 particles (25.6%). This subset of refined particles was then re-extracted and re-centred before another round of 3D auto refinement was carried out. Sample sizes used were sufficient to obtain structures at the reported resolution.
Data exclusions	CryoEM data were processed in Relion. Particles that categorised into poorly defined classes were rejected, while the remaining 292,242 (69.2%) particles were used for further processing. One stable 3D class at a resolution of 5.38 Å was selected for high resolution 3D auto-refinement; this class accounted for a subset of 108,560 particles (25.6%).
Replication	For activity assays, all measurements were performed in triplicate.
Randomization	Not relevant to this study
Blinding	Not relevant to this study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Structures of a RAG-like transposase during cut-and-paste transposition

https://doi.org/10.1038/s41586-019-1753-7

Received: 19 July 2019

Accepted: 30 September 2019

Published online: 13 November 2019

Chang Liu[1,4], Yang Yang[2,3,4] & David G. Schatz[1]*

Transposons have had a pivotal role in genome evolution[1] and are believed to be the evolutionary progenitors of the RAG1–RAG2 recombinase[2], an essential component of the adaptive immune system in jawed vertebrates[3]. Here we report one crystal structure and five cryo-electron microscopy structures of Transib[4,5], a RAG1-like transposase from *Helicoverpa zea*, that capture the entire transposition process from the apo enzyme to the terminal strand transfer complex with transposon ends covalently joined to target DNA, at resolutions of 3.0–4.6 Å. These structures reveal a butterfly-shaped complex that undergoes two cycles of marked conformational changes in which the 'wings' of the transposase unfurl to bind substrate DNA, close to execute cleavage, open to release the flanking DNA and close again to capture and attack target DNA. Transib possesses unique structural elements that compensate for the absence of a RAG2 partner, including a loop that interacts with the transposition target site and an accordion-like C-terminal tail that elongates and contracts to help to control the opening and closing of the enzyme and assembly of the active site. Our findings reveal the detailed reaction pathway of a eukaryotic cut-and-paste transposase and illuminate some of the earliest steps in the evolution of the RAG recombinase.

Transposons are present in all kingdoms of life and move within or between genomes using transposon-encoded transposases[6]. Many DNA transposases and retroviral integrases contain a conserved RNase H-like (RNH) domain that uses three acidic residues (the DDE/D motif) to coordinate magnesium and catalyse DNA cleavage and integration[7]. The RAG1–RAG2 recombinase (RAG), which shares this RNase H catalytic domain[8], generates DNA double-strand breaks at recombination signal sequences (RSSs) to initiate V(D)J recombination in developing lymphocytes of jawed vertebrates[3,9]. The RAG1 catalytic core and RSSs are thought to have evolved from the transposase and terminal inverted repeats (TIRs), respectively, of an ancient *Transib* transposon[10]. Acquisition of a RAG2-like gene by a *Transib* element is proposed to have generated a 'RAG transposon' that subsequently had a key role in the evolution of RAG1–RAG2 loci and V(D)J recombination[2]. Unlike cut-and-paste transposition, which is an excision-and-integration reaction, V(D)J recombination is an excision-and-end joining reaction that rejoins the ends of the excised segment to protect the genome against hazardous insertions (Fig. 1a). Thus, RAG has been subject to different evolutionary constraints to its transposase ancestors, particularly in the events that occur after DNA cleavage.

Transib from *H. zea* (hereafter designated *Transib* unless otherwise specified) is an active transposon with a TIR that resembles a portion of the RSS[4] (Fig. 1b) and a transposase (Transib protein) that cleaves DNA using a nick-hairpin mechanism similar to that of RAG and the hAT family transposase Hermes[5,11] (Fig. 1a). Transib[5], Hermes[12] and RAG[13,14] are active in vitro for the subsequent strand-transfer reaction that completes transposition; however, for RAG, this step is strongly suppressed in vivo[2].

Recent advances in RAG structural biology have clarified the molecular basis for RSS recognition and cleavage[8,15–17]. However, transposition mediated by DDE/D family enzymes that proceed via hairpinning is less well understood, particularly at the final step of integration into target DNA. In contrast to the availability of structures capturing the strand-transfer complexes of bacteriophage Mu[18] and retroviral integrases[19–22], transposon integration has been visualized structurally for only one eukaryotic DNA transposase, Mos1[23,24], which has a catalytic mechanism that does not involve a hairpin intermediate[25]. As the only known active Transib transposase, *H. zea* Transib provides a unique opportunity for the analysis of a RAG2-independent RAG1-family protein and for comparative insights into the effect of RAG2 on RAG1 function and RAG evolution.

Here we describe near-atomic resolution crystal and cryo-electron microscopy (cryo-EM) structures of *H. zea* Transib in the apo form and complexed with intact TIR substrate, nicked TIR substrate, cleaved transposon ends and transposon ends covalently joined to target DNA (Extended Data Figs. 1–4, Extended Data Tables 1, 2). An additional complex, with Transib bound to transposon ends and target DNA before strand transfer, was also observed in and modelled from the cryo-EM data. These structures represent, to our knowledge, the most complete structural description to date of a eukaryotic cut-and-paste transposition reaction, explain the target site sequence preferences of RAG-family transposases and reveal the conformational changes that

[1]Department of Immunobiology, Yale School of Medicine, New Haven, CT, USA. [2]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT, USA. [3]Howard Hughes Medical Institute, Yale University, New Haven, CT, USA. [4]These authors contributed equally: Chang Liu, Yang Yang. *e-mail: david.schatz@yale.edu

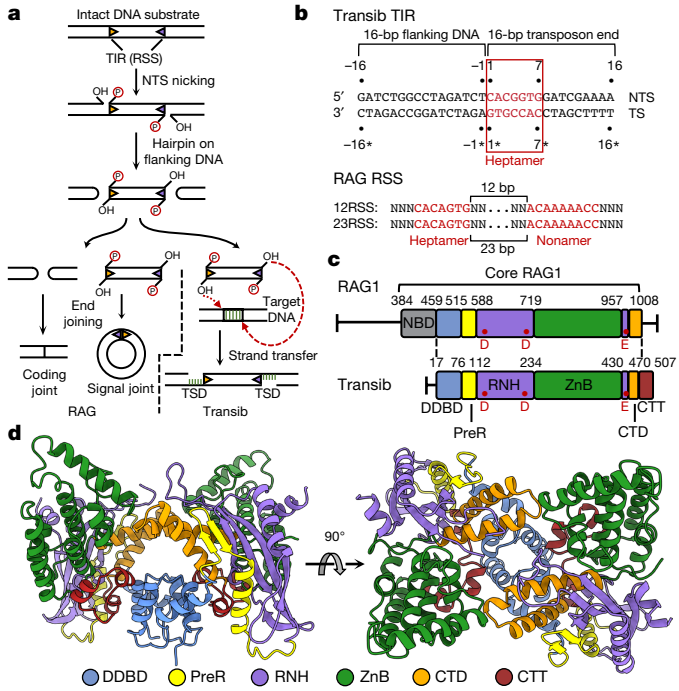


Fig. 1 | Functional and structural overview of *H. zea* Transib. a, Schematic of DNA recombination and transposition pathways of RAG and Transib. RSS or TIR are shown as triangles with wide side indicating heptamer sequence. TSD, target site duplication. **b**, Sequence and numbering of the Transib TIR substrate. Heptamer and nonamer sequences of TIR and RSS are shown in red. TS, transferred strand. NTS, non-transferred strand. The nicking site on Transib TIR is between T−1 and C1 on NTS. **c**, Domain organization of *H. zea* Transib compared with mouse RAG1. Domain boundaries are shown by residue number. Active site carboxylates are labelled in red. **d**, Front and top view of the apo Transib dimer crystal structure.

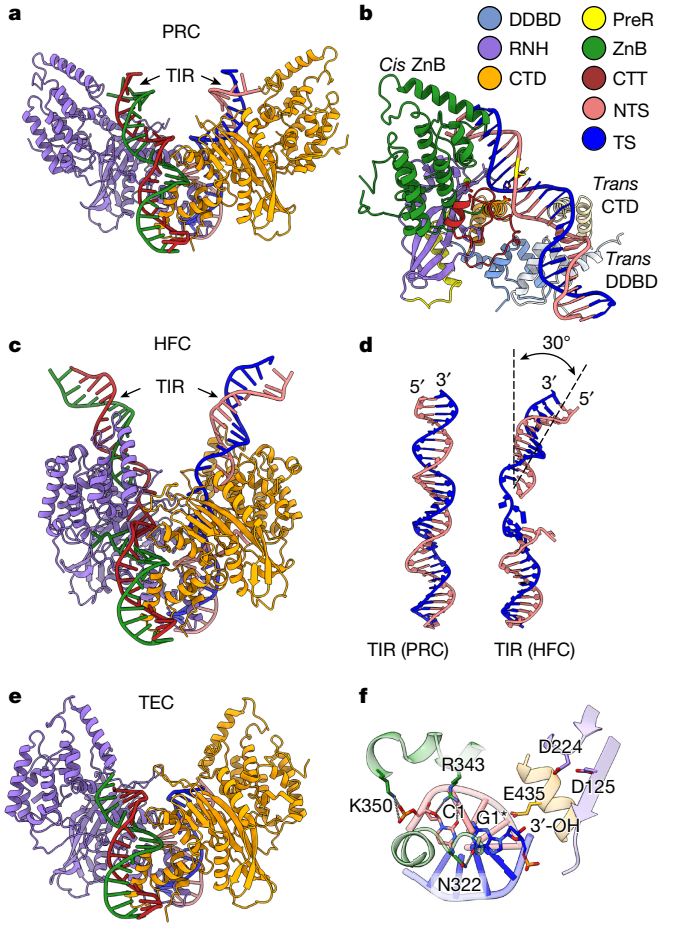


Fig. 2 | Structures of Transib–TIR complexes during transposon binding and excision. a, Overall cryo-EM structure of Transib PRC with intact TIR substrates. Two Transib subunits are coloured in orange and purple. **b**, *Trans* architecture of Transib–TIR complex. DDBD and CTD from *trans* Transib are in pale shades. Mg^{2+} ion, green sphere; catalytic carboxylates, red sticks; scissile phosphate is highlighted in yellow. **c**, Overall cryo-EM structure of Transib HFC with nicked TIR substrates. **d**, Comparison of TIR substrates from PRC and HFC. **e**, Overall cryo-EM structure of Transib TEC with catalytically cleaved transposon end DNAs. **f**, Transposon end nucleotides are stabilized by ZnB domain residues, but the 3′-OH is not coordinated for the subsequent strand transfer reaction.

enable the same catalytic centre to perform both transposon excision and integration.

Opening of Transib upon TIR engagement

Apo Transib exhibits a modular domain arrangement similar to that of RAG1[8] (Fig. 1c, d). The N-terminal dimerization and DNA-binding domain (DDBD) serves as the dimerization interface and is connected by an extended pre-RNase H (PreR) loop to a split RNH domain containing three conserved catalytic carboxylates[5] (D125, D224 and E435), all of which are required for activity[5] (Extended Data Fig. 1c, d). E435 is separated from the rest of RNH by two zinc-binding domains, ZnC_2 and ZnH_2 (collectively, ZnB), which form a C_2H_2 zinc finger (Extended Data Fig. 5a), as in RAG1[8]. The following C-terminal domain (CTD) folds back to interact with DDBD, and the protein ends with a C-terminal tail (CTT) of about 30 amino acids, made up of three short helices that bridge from DDBD to ZnB. The absence of a nonamer-binding domain (NBD) (Extended Data Fig. 5b) is consistent with the observation that Transib TIRs have sequence similarity to the heptamer but not the nonamer of the RSS[4,26] (Fig. 1b).

Despite low (16.4%) sequence identity between the Transib and RAG1 core, individual domains from the two proteins are readily superimposable (Extended Data Fig. 5a), providing support for the model in which Transib and RAG1 are evolutionarily related. These alignments also reveal several differences between Transib and RAG1, three of which (red boxes in Extended Data Fig. 5a) are extended structural elements in RAG1, absent from Transib, that together constitute a substantial portion of the RAG2-binding interface in RAG1 (Extended Data Fig. 5c). These three missing elements explain the absence of a RAG2-like entity in Transib and poor RAG2 binding by Transib in vitro[4,26].

Binding of intact TIR substrate to form the pre-reaction complex (PRC) induces a marked relocation of the ZnB domains, from being tightly packed components of the enzyme core to lateral extensions that jut away from the core (Fig. 2a, Extended Data Fig. 6a and Supplementary Video 1). This 49° rotation and 26 Å centroid movement of ZnB exposes the TIR-binding grooves (Extended Data Fig. 6a) and is twice as large as the RAG1 ZnB-domain movement that occurs on intact RSS binding[16]. Viewed from the front, the Transib PRC resembles a butterfly with wings spread and DNA as antennae, with ZnB domain rotation constituting an 'unfurling' of the wings, one from the back and the other from the front of the butterfly (Fig. 2a, Extended Data Fig. 6a).

The Transib PRC adopts a *trans* architecture in which each TIR engages the active site of one Transib (the *cis* subunit) but is bound primarily by the other Transib (the *trans* subunit) (Fig. 2b), similar to RAG and other DDE transposases and retroviral integrases[7,15,16]. CTT from the *cis* subunit tracks through the heptamer major groove and interacts with the backbone of TIR position 3 (Extended Data Fig. 6b). *Trans* DNA binding interactions include base-specific interactions between CTD and TIR positions 5–7 and between DDBD and the phosphate backbone at TIR positions 8–13 (Extended Data Fig. 6c, d). No interaction is observed beyond position 13, and consistent with this,

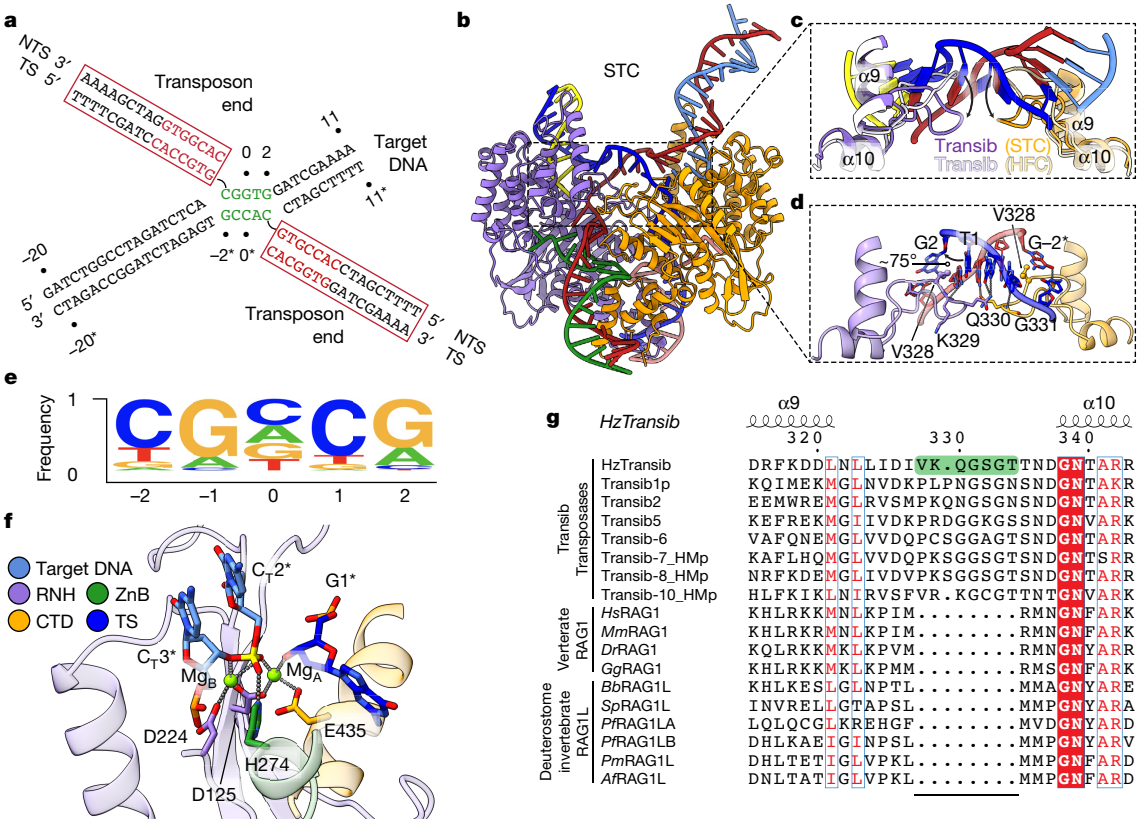


Fig. 3 | Transposon end integration and strand transfer complex. **a**, Schematic of strand transfer product. Heptamer and target site sequences are coloured red and green, respectively. **b**, Overall cryo-EM structure of Transib in complex with naturally generated strand transfer product. **c**, Different conformations of α9–α10 target site-binding loop in STC and HFC. **d**, Interactions between the α9–α10 loop and the 5-bp target site. Hydrogen bonds are shown as dashed lines. **e**, Sequence logo representing nucleotide frequencies at Transib TIR-integration sites. **f**, Active site of the TCC model. Mg^{2+} ions, green spheres. Nucleotide residues in target DNA are indicated with subscript T. **g**, Sequence alignment of Transib transposases, vertebrate RAG1 and deuterostome invertebrate RAG1L proteins. Residue numbers and secondary structure annotation are for *H. zea* Transib. The α9–α10 loop in *H. zea* Transib is highlighted in green. *Hs*, *Homo sapiens* (human); *Mm*, *Mus musculus* (mouse); *Dr*, *Danio rerio* (zebrafish); *Gg*, *Gallus gallus* (chicken); *Bb*, *Branchiostoma belcheri* (amphioxus); *Sp*, *Strongylocentrotus purpuratus* (purple sea urchin); *Pf*, *Ptychodera flava* (acorn worm); *Pm*, *Petromyzon marinus* (sea lamprey); and *Af*, *Asterias forbesi* (sea star).

serial 3′ truncations of the TIR demonstrate that cleavage in vitro is robust with only the first 13 bp of the TIR (Extended Data Fig. 1e).

The PRC is a cleavage-incompetent complex in which the scissile phosphate for nicking is far from the active site and E435 is not positioned for catalysis (Extended Data Fig. 6e). This indicates that a substantial structural alteration will be required before nicking of the NTS could take place.

Transib closure accompanies catalysis

Incubation of Transib with nicked TIR substrate at 30 °C in Ca^{2+} yielded a complex that is poised for hairpin formation, referred to as the hairpin-forming complex (HFC). The HFC is more compact than the PRC, with the ZnB domains having undergone a major 51° inward rotation along an axis nearly perpendicular to that of the original outward movement (Fig. 2c, Extended Data Fig. 6g and Supplementary Video 1). This inward folding of the ZnB wings is accompanied by several other changes in the complex. First, flanking DNA is rotated about 180° and tilted around 30° towards the *cis* ZnB domain, with bases C1 and A-1* becoming flipped out of the helix (Fig. 2d, Extended Data Fig. 6h–j). A similar DNA rotation is seen in RAG-nicked RSS structures[15,16]. Second, an approximately 6 Å movement of E435 has led to full assembly of the active site (Extended Data Fig. 6k). Third, HFC formation results in numerous new *cis* Transib–DNA contacts. The first 3 bp of the heptamer make extensive base-specific interactions with helices α10 and α16 of the *cis* subunit (Extended Data Fig. 6h) and the extrahelical C1 base is buried in a pocket formed by helices α10 and α12 and a loop of CTT (Extended Data Fig. 6i). ZnB enfolds the flanking DNA (Fig. 2c) and interacts with the first 7 bp of flanking DNA; in the PRC, such interactions extended only to position −4 (Extended Data Fig. 6l, m). Owing to its lack of a RAG2 subunit, interactions of Transib with flanking DNA are much less extensive than for RAG, in which RAG2 contacts extend to position −15 in the PRC and HFC[15,16].

Transib reopens upon DNA cleavage

The *Transib* transposon end complex (TEC) structure, in which hairpin formation and release of flanking DNA has occurred, provides a view of post-cleavage events for hAT/RAG family enzymes. Release of the flanking DNA hairpin ends is associated with a 26° rotation of the ZnB domains that partially spreads the wings of the complex (Fig. 2e, Extended Data Fig. 6n and Supplementary Video 1). C1 of the heptamer is switched from its flipped-out position to base pair with G1*, and transposon end DNA has become largely superimposable with that in the PRC (Extended Data Fig. 6o). In the absence of flanking DNA, the ZnB domains are able to tilt and interact with the exposed heptamer ends, physically sequestering them through interactions involving N322, R343 and K350 (Fig. 2f). The 3′-OH that will be the nucleophile for the target-integration reaction is not in close proximity to the three active site carboxylates (Fig. 2f), indicating that substantial distortions of the transposon end and conformational changes in Transib will be necessary for the strand transfer reaction.

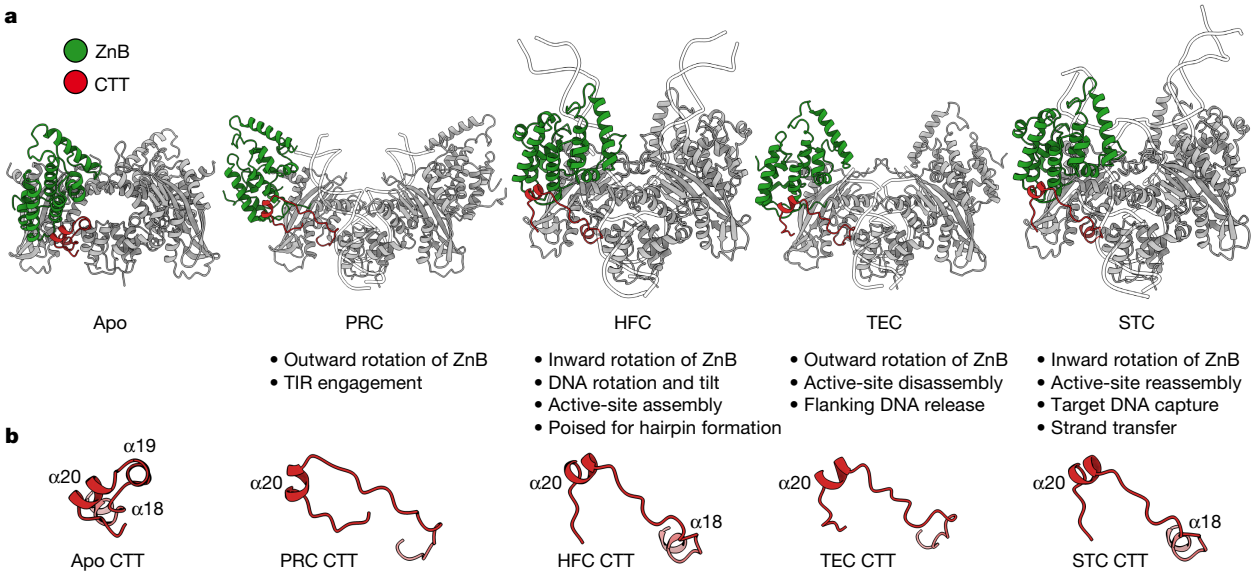


Fig. 4 | Transib CTT conformational changes during transposition. a, **b**, Side-by-side comparison of five *H. zea* Transib structures, with ZnB and CTT domains coloured in green and red, respectively.

The TEC structure illustrates how Transib prepares for target capture and reveals several structural differences from RAG and other transposases. The substantial outward rotation of ZnB seen in the TEC exposes the DNA-binding groove and probably facilitates flanking DNA release and target capture. No such movement is seen for RAG or Hermes[12,15]. In addition, the interactions of Transib with the cleaved transposon ends might shield the DNA from DNA repair enzymes and inhibit end joining. Similar interactions with the cleaved RSS ends are not seen in the RAG signal end complex (SEC)[15], a difference that might reflect the different evolutionary constraints faced by Transib and RAG. Finally, dislocation of the 3′-OH nucleophile out of the Transib active site in the TEC is not observed in the RAG SEC or TEC of other transposases[12,15,23,27].

Target DNA capture and strand transfer

During transposition, hairpin formation and flanking DNA release are followed by non-covalent capture of target DNA to form the target-capture complex (TCC) and then by the strand transfer reaction that covalently joins the transposon ends to target DNA to form the strand transfer complex (STC). The Transib TCC and STC were formed through cleavage of intact TIR substrates without the provision of a specific target DNA. One 3D class of Transib–TIR complexes contained a clearly resolved density connecting the catalytic centres of the two Transib subunits (Extended Data Fig. 7a), which was determined (see Methods) to be the 5-bp target site generated after attack of the transposon ends at a 5′-CGGTG-3′ sequence in an additional TIR substrate molecule (Fig. 3a).

The STC structure reveals that engagement of target DNA triggers active site reassembly driven by rotational closure of the ZnB domains, which now enfold target DNA in much the same manner that they previously bound flanking DNA in the HFC (Fig. 3b, Extended Data Fig. 7b and Supplementary Video 1). The α9–α10 loop has moved downward towards the RNH domain (Fig. 3c) and interacts extensively with target site DNA (Fig. 3d). Target site DNA exhibits sharp (approximately 75°) bends 1 bp from each end, resulting in an overall directional change of about 150° (Fig. 3b, Extended Data Fig. 7c). V328 fills the gaps left by the breaks in base-stacking on the continuous strands, stabilizing the highly kinked DNA conformation (Fig. 3d).

Transib exhibits a 5′-CGNCG-3′ transposition target site consensus sequence and target sites almost always contain a 5′-YR-3′ dinucleotide step at one or both ends (Fig. 3e and Supplementary Table 1). This preference is probably because of the inherent deformability and reduced base-stacking of a pyrimidine–purine step[28]. Notably, the GC-rich Transib target site DNA remains fully base-paired in the STC despite its highly distorted duplex structure.

A trifurcation of density observed at the transposon end–target DNA junction suggests that the cryo-EM map represents a mixture of Transib in complex with target DNA before (TCC) and after (STC) transposon end integration (Extended Data Fig. 7d). Indeed, calculation of the difference map between the cryo-EM reconstruction and the STC model suggested that a proportion of the particles contain uncleaved target DNA (Extended Data Fig. 7e) and enabled modelling of intact target DNA in the cryo-EM density. In this TCC model, the active site captures two Mg^{2+} ions (Fig. 3f), whereas none are observed in the disassembled active site of the TEC (Fig. 2f). One non-bridging oxygen is hydrogen bonded with H274 (Fig. 3f). This histidine, which is conserved in several eukaryotic transposase superfamilies[29,30], has been proposed to be a key component of a DDHE/D (as opposed to DDE/D) enzyme active site[30], and our data are consistent with this proposal. The distances separating the scissile phosphate and the attacking oxygen and the two metal ions in the TCC model strongly suggest that the active site could catalyse the strand-transfer reaction[31] (Extended Data Fig. 7f).

CTT helps to drive Transib domain closure

During the two cycles of Transib opening and closing, CTT acts as an accordion-like element that extends and refolds in concert with the unfurling and furling of the wings of Transib (Fig. 4a, b and Supplementary Video 1). In apo Transib, CTT is a compact bundle of three short helices, α18–α20 (Fig. 4b). α20 is anchored through interactions with helices α12 and α13 of ZnB and stays almost static relative to ZnB throughout the transposition cycle (Extended Data Fig. 8a, b). By contrast, helices α18 and α19 markedly alter their secondary structures during the structural rearrangements of Transib. The large rotation of ZnB that accompanies binding of intact TIR DNA substantially elongates and deforms helices α18 and α19 (Fig. 4a, b). This CTT coil might help to drive the inward movement of ZnB and closure of the Transib dimer in the subsequent PRC-to-HFC transition, during which helix α18 reforms (Fig. 4a, b). Helix α18 becomes deformed again during Transib opening and flanking DNA release in the TEC and then reforms during Transib closure and target DNA engagement in the STC (Fig. 4a, b). Helix α18 is particularly well conserved across Transib

proteins, and the hydrophobic residues that anchor helix α20 to ZnB also exhibit sequence conservation (Extended Data Fig. 8e). Hence, CTT is probably an ancient and functionally conserved component of many Transib transposases, and its deletion from Transib almost abolished DNA cleavage activity (Extended Data Fig. 8c).

The C-terminal tails of jawed vertebrate RAG1 and invertebrate RAG1-like (RAG1L) proteins, including the *Branchiostoma belcheri* RAG1L subunit of the ProtoRAG transposase from amphioxus[32], show sequence similarity only with helix α18 of *H. zea* Transib CTT (Extended Data Fig. 8e) and are unlikely to perform functions similar to that of CTT. The RAG1 C-terminal tail is dispensable for activity, and the functionally important *B. belcheri* RAG1L C-terminal tail interacts with TIR DNA downstream of the heptamer and not with ZnB, and shares no structural similarity with *H. zea* Transib CTT[33] (Extended Data Fig. 8d). Hence, the CTT module of RAG1 family proteins has apparently been readily adapted during evolution to address different functional imperatives.

RAG2 acquisition and transposase evolution

The lack of structural information for Transib has made it difficult to explore the structural and functional implications of the acquisition of a RAG2-like subunit by RAG1 early in evolution. The absence of RAG2 is likely to be of particular relevance for the large domain excursions that characterize the Transib transposition reaction (Fig. 4a). The outward rotation of ZnB that accompanies initial DNA binding in the PRC provides extensive access to DNA-binding surfaces, thereby helping to compensate for the lack of stabilizing RAG2-flanking DNA interactions[15,16]. The subsequent domain closure that yields the Transib HFC creates ZnB-flanking DNA interactions (Extended Data Fig. 6m) that are contributed predominantly by RAG2 in the RAG HFC[15,16]. Interdimer interactions mediated by RAG2 stabilize the closed configuration of the RAG HFC[15] and their absence might help to explain the need for a unique CTT to help to drive inward rotation during Transib HFC formation.

Perhaps most notably, the α9–α10 target site-interaction loop of *H. zea* Transib (Fig. 3c, d), a nearly ubiquitous feature of predicted Transib proteins, is absent in RAG1 and invertebrate RAG1-like proteins predicted to have a RAG2-like partner (Fig. 3g). By stabilizing target DNA in the TCC and STC, the Transib target site-interaction loop probably compensates for the absence of stabilizing RAG2–DNA interactions. We propose that acquisition of a RAG2-like gene by a *Transib* transposon to give rise to the first RAG1–RAG2 transposon[2] set in motion two linked evolutionary processes in RAG1: acquisition of new RAG2-binding interfaces (Extended Data Figs. 5a, 7g) and loss of the target site-interaction loop, which was now no longer needed for the stabilization of target DNA.

The structure of the *H. zea* Transib STC reveals distinctive structural and mechanistic features of cut-and-paste transposition. The large overall target DNA distortion created by the deep binding pocket of Transib contrasts with the relatively mild target DNA bend and flat target DNA-binding groove in retroviral integrases[19–22] (Extended Data Fig. 7c). A second distinctive feature of Transib is the large protein conformational change that occurs during target DNA capture (Fig. 4a and Supplementary Video 1). By contrast, Mos1[23,24] and retroviral integrases[19,21,34] adopt very similar structures before and after target DNA capture. Finally, the Transib STC structure helps to explain multiple features of RAG-family transposition: the preferred 5-bp target site duplication length, GC-rich target sites[5,10,13,32], target site hotspot sequence preferences[13,35] and the ability of mismatches and other DNA distortions to stimulate transposition by RAG[35,36].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1753-7.

1. Feschotte, C. & Pritham, E. J. DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.* **41**, 331–368 (2007).
2. Carmona, L. M. & Schatz, D. G. New insights into the evolutionary origins of the recombination-activating gene proteins and V(D)J recombination. *FEBS J.* **284**, 1590–1605 (2017).
3. Gellert, M. V(D)J recombination: RAG proteins, repair factors, and regulation. *Annu. Rev. Biochem.* **71**, 101–132 (2002).
4. Chen, S. & Li, X. Molecular characterization of the first intact Transib transposon from *Helicoverpa zea*. *Gene* **408**, 51–63 (2008).
5. Hencken, C. G., Li, X. & Craig, N. L. Functional characterization of an active Rag-like transposase. *Nat. Struct. Mol. Biol.* **19**, 834–836 (2012).
6. Craig, N. L. in *Mobile DNA III* (eds Craig, N. L. et al.) 3–39 (ASM Press, 2015).
7. Montaño, S. P. & Rice, P. A. Moving DNA around: DNA transposition and retroviral integration. *Curr. Opin. Struct. Biol.* **21**, 370–378 (2011).
8. Kim, M. S., Lapkouski, M., Yang, W. & Gellert, M. Crystal structure of the V(D)J recombinase RAG1–RAG2. *Nature* **518**, 507–511 (2015).
9. Schatz, D. G. & Swanson, P. C. V. V(D)J recombination: mechanisms of initiation. *Annu. Rev. Genet.* **45**, 167–202 (2011).
10. Kapitonov, V. V. & Jurka, J. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol.* **3**, e181 (2005).
11. Zhou, L. et al. Transposition of hAT elements links transposable elements and V(D)J recombination. *Nature* **432**, 995–1001 (2004).
12. Hickman, A. B. et al. Structural basis of hAT transposon end recognition by Hermes, an octameric DNA transposase from *Musca domestica*. *Cell* **158**, 353–367 (2014).
13. Agrawal, A., Eastman, Q. M. & Schatz, D. G. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* **394**, 744–751 (1998).
14. Hiom, K., Melek, M. & Gellert, M. DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. *Cell* **94**, 463–470 (1998).
15. Ru, H. et al. Molecular mechanism of V(D)J recombination from synaptic RAG1–RAG2 complex structures. *Cell* **163**, 1138–1152 (2015).
16. Kim, M. S. et al. Cracking the DNA code for V(D)J recombination. *Mol. Cell* **70**, 358–370 (2018).
17. Ru, H. et al. DNA melting initiates the RAG catalytic pathway. *Nat. Struct. Mol. Biol.* **25**, 732–742 (2018).
18. Montaño, S. P., Pigli, Y. Z. & Rice, P. A. The μ transpososome structure sheds light on DDE recombinase evolution. *Nature* **491**, 413–417 (2012).
19. Maertens, G. N., Hare, S. & Cherepanov, P. The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature* **468**, 326–329 (2010).
20. Yin, Z. et al. Crystal structure of the Rous sarcoma virus intasome. *Nature* **530**, 362–366 (2016).
21. Ballandras-Colas, A. et al. A supramolecular assembly mediates lentiviral DNA integration. *Science* **355**, 93–95 (2017).
22. Passos, D. O. et al. Cryo-EM structures and atomic model of the HIV-1 strand transfer complex intasome. *Science* **355**, 89–92 (2017).
23. Richardson, J. M., Colloms, S. D., Finnegan, D. J. & Walkinshaw, M. D. Molecular architecture of the Mos1 paired-end complex: the structural basis of DNA transposition in a eukaryote. *Cell* **138**, 1096–1108 (2009).
24. Morris, E. R., Grey, H., McKenzie, G., Jones, A. C. & Richardson, J. M. A bend, flip and trap mechanism for transposon integration. *eLife* **5**, e15537 (2016).
25. Dawson, A. & Finnegan, D. J. Excision of the *Drosophila* mariner transposon Mos1. Comparison with bacterial transposition and V(D)J recombination. *Mol. Cell* **11**, 225–235 (2003).
26. Carmona, L. M., Fugmann, S. D. & Schatz, D. G. Collaboration of RAG2 with RAG1-like proteins during the evolution of V(D)J recombination. *Genes Dev.* **30**, 909–917 (2016).
27. Davies, D. R., Goryshin, I. Y., Reznikoff, W. S. & Rayment, I. Three-dimensional structure of the Tn5 synaptic complex transposition intermediate. *Science* **289**, 77–85 (2000).
28. Lankaš, F., Sponer, J., Langowski, J. & Cheatham, T. E. III. DNA basepair step deformability inferred from molecular dynamics simulations. *Biophys. J.* **85**, 2872–2883 (2003).
29. Yuan, Y. W. & Wessler, S. R. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc. Natl Acad. Sci. USA* **108**, 7884–7889 (2011).
30. Hickman, A. B. et al. Structural insights into the mechanism of double strand break formation by Hermes, a hAT family eukaryotic DNA transposase. *Nucleic Acids Res.* **46**, 10286–10301 (2018).
31. Yang, W., Lee, J. Y. & Nowotny, M. Making and breaking nucleic acids: two-Mg^{2+}-ion catalysis and substrate specificity. *Mol. Cell* **22**, 5–13 (2006).
32. Huang, S. et al. Discovery of an active RAG transposon illuminates the origins of V(D)J recombination. *Cell* **166**, 102–114 (2016).
33. Zhang, Y. et al. Transposon molecular domestication and the evolution of the RAG recombinase. *Nature* **569**, 79–84 (2019).
34. Hare, S., Gupta, S. S., Valkov, E., Engelman, A. & Cherepanov, P. Retroviral intasome assembly and inhibition of DNA strand transfer. *Nature* **464**, 232–236 (2010).
35. Tsai, C. L., Chatterji, M. & Schatz, D. G. DNA mismatches and GC-rich motifs target transposition by the RAG1/RAG2 transposase. *Nucleic Acids Res.* **31**, 6180–6190 (2003).
36. Lee, G. S., Neiditch, M. B., Sinden, R. R. & Roth, D. B. Targeted transposition by the V(D)J recombinase. *Mol. Cell. Biol.* **22**, 2068–2077 (2002).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Cloning of *H. zea* Transib transposase and substrates

The full-length or an N-terminal truncated fragment (residues 17–507) of *H. zea* Transib transposase fused to a C-terminal His_6 tag or an N-terminal maltose-binding protein (MBP) tag were cloned into pFastBac1 expression vector (ThermoFisher Scientific) between BamHI and HindIII restriction sites. pB-5′/3′TIR, a derivative of pBR322 containing the TIR substrate for ProtoRAG transposases, was described previously[33]. To generate TIR substrate for Transib transposases, the ProtoRAG 5′TIR and 3′TIR of pB-5′/3′TIR were substituted by the first 51 bp and 50 bp of *Transib* transposon 5′TIR and 3′TIR, respectively, using In-Fusion cloning (Clontech). The PCR-amplified and linearized Transib substrate contains a *Transib* 5′TIR and 3′TIR separated by 411 bp between their tips, 126 bp of DNA flanking 5′TIR and 276 bp of DNA flanking 3′TIR. The whole substrate was depleted of 5′-CAC-3′ sequence instances except for those contained in 5′TIR and 3′TIR regions.

Protein expression and purification

MBP- or His_6-tagged Transib transposase was expressed in Sf9 insect cells using the Bac-to-Bac Baculovirus Expression System (ThermoFisher Scientific) according to the manufacturer's protocol. Cells expressing His_6-tagged Transib transposase were resuspended in lysis buffer (20 mM Tris-HCl, pH7.5, 500 mM NaCl, 1 mM dithiothreitol (DTT)) and lysed by six passes through an Emulsiflex C3 homogenizer (Avestin). Cell lysate was cleared by centrifugation at 40,000 r.p.m. (~146,000*g*) using a Type 50.2 Ti rotor (Beckman Coulter) for 1 h at 4 °C and was mixed with pre-equilibrated Ni-NTA Agarose resin (Qiagen) for 2 h with continual rotation. The resin was loaded onto a gravity flow column, washed with 5× column volume (CV) of lysis buffer and protein eluted with 5× CV of elution buffer (20 mM Tris-HCl, pH7.5, 200 mM NaCl, 20 mM imidazole, 1 mM DTT). The eluate was further purified and buffer exchanged using a Superdex 200 Increase 10/300 GL size-exclusion chromatography column (GE Healthcare) in 20 mM Tris-HCl, pH 7.5, 200 mM NaCl and 1 mM Tris(2-carboxyethyl) phosphine hydrochloride (TCEP-HCl). Cells expressing MBP-tagged Transib transposase were resuspended in lysis buffer (20 mM Tris-HCl, pH7.5, 500 mM NaCl, 1 mM DTT) and purified using amylose resin (New England BioLabs) in 20 mM Tris-HCl, pH7.5, 200 mM NaCl, 1 mM DTT, followed by size-exclusion chromatography purification in 20 mM Tris-HCl, pH 7.5, 200 mM NaCl and 1 mM TCEP. Both forms of Transib protein are a dimer in solution and show TIR-dependent nuclease activity (Extended Data Fig. 1).

Mutant Transib proteins with active site mutations or C-terminal tail (CTT) truncation (removal of residues 478–507) were fused to an MBP tag and purified in the same way as MBP-tagged wild-type Transib transposase.

His_6-tagged human HMGB1 with C-terminal truncation (residues 1–165) was expressed in *Escherichia coli* BL21 (DE3) and purified as previously described[15].

Sf9 cells were obtained from Thermo Fisher Scientific. Cells lines used were not authenticated or tested for mycoplasma contamination.

Crystallization and data collection

Purified His_6-tagged Transib transposase was concentrated to ~6.3 mg ml^{-1} and used in crystallization screening. Transib crystals were grown by sitting-drop vapour diffusion at 20 °C in 100 mM HEPES, pH 7.0, 0.7–0.8 M NaH_2PO_4 and 0.75 M KH_2PO_4. Crystals were cryo-protected in crystallization solution supplemented with 17.5% glycerol and flash frozen in liquid nitrogen. Heavy atom derivatives of Transib crystals were prepared by soaking crystals in cryo-protection solution supplemented with 1 M NaBr for 2–5 min, 0.5 M NaI for 2–5 min, 2.5 mM K_2OsCl_6 for 2 h, 2.5 mM K_2PtCl_4 for 2 h, or 2.5 mM ethylmercury thiosalicylate (EMTS) for 2 h. Data were collected at 100 K at beamline 24ID-E and 24ID-C of the Advanced Photon Source at Argonne National Laboratory. The dataset for the native crystal was collected at 0.9792 Å. The datasets for Br-, I-, Os-, Pt- and Hg-derivative crystals were collected at 0.9197 Å, 1.4586 Å, 1.1398 Å, 1.0718 Å and 1.0087 Å, respectively. All X-ray diffraction data were indexed, integrated and scaled with the XDS package[37] (Extended Data Table 1).

Crystal structure determination and refinement

Phases were determined with native crystal dataset and five heavy-atom-derivative datasets by multiple isomorphous replacement with anomalous scattering (MIRAS) method. Heavy atom sites were identified using SHELXD[38] and the structure was determined using AutoSol[39]. The initial model was built automatically using AutoBuild[40] of PHENIX software package and manually rebuilt in COOT[41]. The model was refined in PHENIX[42] with non-crystallographic symmetry (NCS) restraints. The final structure was refined to 3.0 Å with R_{work} and R_{free} of 22.0% and 27.7%, respectively. Due to poor electron densities, residues 17–20, 235–238, 247–264 and 502–507 were not included in the final model. The structure was validated with MolProbity[43]. 92.98% of residues are in the favoured regions of the Ramachandran plot, 6.47% in additional allowed regions, and 0.56% in the disallowed region.

Transib–TIR complex assembly

The 24-bp intact TIR substrate was generated by annealing equimolar amounts of two complementary oligonucleotides: 5′-CTAGATCTCACGGTG GATCGAAAA-3′ and 5′-TTTTCGATCCACCGTG*AGATCTAG-3′ (heptamer sequence is underlined and the asterisks indicates a phosphorothioate bond introduced between the two nucleotide residues). The 32-bp intact TIR substrate was generated by annealing equimolar amounts of the two oligonucleotides: 5′-GATCTGGCCTAGATCTCACGGTG GATCGAAAA-3′ and 5′-TTTTCGATCCACCGTGAGATCTAGGCCAGATC-3′. 32 bp nicked TIR substrate was generated by annealing equimolar amounts of the following three oligonucleotides: 5′-GATCTGGCCTAGA TCT-3′, 5′-CACGGTGGATCGAAAA-3′ and 5′-TTTTCGATCCACCGTGAGA TCTAGGCCAGATC-3′ (a phosphorothioate bond was introduced between the heptamer and flanking DNA on transferred strand for the nicked TIR substrates used in Transib–TIR complex reconstitution in the present of Mg^{2+}). To reconstitute the Transib–intact TIR complex, purified MBP-tagged Transib was mixed with 24 bp intact TIR substrate and HMGB1 in a 1:2:2 molar ratio in the presence of Mg^{2+} at 4 °C for 1 h, followed by size-exclusion chromatography purification in 20 mM Tris-HCl, pH 7.5, 50 mM KCl, 10 mM $MgCl_2$, 1 mM TCEP. Transib-nicked TIR complex was reconstituted by mixing MBP-tagged Transib, 32 bp nicked TIR substrate and HMGB1 in a 1:2:2 molar ratio in the present of Mg^{2+} at 4 °C or in the presence of Ca^{2+} at 30 °C for 1 h, followed by size-exclusion chromatography purification. Catalytically active Transib-TIR complex was reconstituted by mixing MBP-tagged Transib with 32 bp intact TIR substrate and HMGB1 in a 1:2:2 molar ratio in the presence of Mg^{2+}, and was allowed to react at 30 °C for 50 min before being frozen on cryo-EM grids.

Cryo-EM sample preparation and data acquisition

Purified Transib–TIR complex (3.5 µl at ~1.2 µM) was applied to freshly glow-discharged Quantifoil 300 mesh or 200 mesh holey carbon grids with R 1.2/1.3 hole pattern (Electron Microscopy Sciences). Grids were blotted for 5.5 s under 100% humidity and plunge-frozen in liquid nitrogen-cooled liquid ethane using a Vitrobot Mark IV (ThermoFisher Scientific). Cryo-EM datasets were collected on a Titan Krios G2 electron microscope (Yale University) operated at 300 kV equipped with a GIF Quantum LS imaging filter (Gatan) and a K2 summit direct electron detector (Gatan) in super-resolution mode. The image stacks were collected at a nominal magnification of 130,000×, corresponding to 0.525 Å per super-resolution pixel, at a dose rate of 7.0–7.5 e^- per physical pixel

per s. The total exposure time for each movie was 8 s, thus leading to a total accumulated dose of 50.8–54.4 e^{-} Å^{-2}, which was fractionated into 40 frames. All movies were recorded with a defocus ranging from −1.5 to −2.5 µm. The statistics of cryo-EM data acquisition are summarized in Extended Data Table 2.

Image processing

Dose-fractionated super-resolution movies were binned over 2 × 2 pixels, yielding a pixel size of 1.05 Å, then subjected to motion correction and dose-weighting using MotionCorr2[44]. The non-dose-weighted aligned images were used for contract transfer function estimation by CTFFIND-4.1.10[45]. The dose-weighted images were used for autopicking, classification and reconstruction. For Transib–TIR complex datasets, roughly 40,000 particles were automatically picked using a Laplacian-of-Gaussian blob detection in RELION-3.0[46], followed by a round of 2D classification to generate templates for a new round of autopicking. The newly autopicked particles were subjected to multiple rounds of 2D classification in RELION-3.0 to remove junk particles. Particles in good 2D classes were extracted for initial model generation in RELION-3.0. The initial model was low-pass filtered to 50 Å to serve as a starting reference for 3D auto-refinement in RELION-3.0 using all particles in good 2D classes. The signal corresponding to MBP regions was then subtracted, followed by 3D classification with a mask encompassing the Transib transposases dimer plus TIRs DNA region. Good 3D classes were selected and iteratively refined to yield high-resolution maps in RELION-3.0 with either C1 or C2 symmetry. To improve the map quality and interpretability of the Transib ZnB domains in Transib-TIR PRC, the particles from good 3D class(es) were symmetry-expanded and subjected to masked 3D classification with residual signal subtraction focusing on the Transib ZnB domain using a previously published procedure[47]. All refinements followed the gold-standard procedure, in which two half datasets were refined independently. The overall resolution was estimated based on the Fourier shell correlation (FSC) cut-off at 0.143 between the two half-maps, after a soft mask was applied to mask out solvent region. The final maps were sharpened within RELION-3.0. Local resolution variation was estimated from the two half-maps using ResMap[48].

Cryo-EM model building and refinement

The crystal structure of Transib dimer was rigid-fitted into the Transib–TIR complexes cryo-EM maps in UCSF Chimera[49]. Owing to large domain movements, the Transib ZnB domains were fitted separately from the other part of the structural model. The DNA fragments corresponding to heptamer plus the first 16 bp of coding flank from RAG–RSS PRC (PDB 6CIK) or HFC (PDB 5ZE0) structures were first fitted into the Transib PRC or HFC cryo-EM map, respectively, and mutated to the input TIR sequence in COOT. The complex resulting from incubation of Transib with nicked TIR substrate at 4 °C in the presence of Mg^{2+} adopted a catalytically incompetent conformation very similar to that of the Transib-intact TIR complex (Extended Data Fig. 6f). This complex is referred to as the PRC with nicked TIRs. For Transib STC structure, the modelling and sequence registers of the target DNA are based on the following observations. (1) The well-defined cryo-EM density for the target site suggests a 5′-YRRYR-3′ motif (Y, pyrimidine; R, purine) (Extended Data Fig. 7a), 5′-CGGTG-3′ is the only match throughout the entire sequence of the input TIR substrate DNA. (2) Transib prefers GC-rich target site for integration. In vitro transposition has shown that *Hz*Transib can mediate transposon integration at target sites with an exact 5′-CGGTG-3′ sequence (Supplementary Table 1). (3) Reconstruction of *Hz*Transib STC cryo-EM map without imposing C2 symmetry shows asymmetric DNA helix density at two flanking DNA-binding regions in the *Hz*Transib dimer. The cryo-EM density for two flanking site DNA helices exhibits a 7–9 bp difference in length, which coincides with 18-bp and 9-bp flanking DNA on two sides of the 5′-CGGTG-3′ sequence in our TIR DNA substrate. By contrast, reconstructing the cryo-EM map of other *Hz*Transib–TIR complexes without C2 symmetry results in a map with nearly perfect two-fold symmetry. (4) The sequence registers for target DNA in this model is also largely supported by the cryo-EM density features. The structural models were manually adjusted and rebuilt in COOT and refined using PHENIX real-space refinement with secondary structure restraints, rotamer restraints, Ramachandran restraints and NCS constraints (except for *Hz*Transib–TIR STC, in which no NCS was applied). The final structures were validated with MolProbity. The final HFC and STC structures contain amino acid residues 21–500 of *Hz*Transib and most TIR DNA nucleotides, except for the two most distal base pairs of the transposon end-flanking DNA and 5′ end of target DNA. In PRC structures, residues 17–20, 131–141 and 245–252 of *Hz*Transib and the most distal base pair of the transposon end-flanking DNA are not modelled owing to poor density. The TEC model contains all 16 bp of transposon end DNA. *Hz*Transib residues 17–20, 136–141 and 245–265 are disordered and are not included in the final TEC model. No HMGB1 density was seen in any of the cryo-EM density maps, and thus was not included in the cryo-EM atomic models. All molecular representations were generated in UCSF Chimera and UCSF ChimeraX[50]. Sequence alignments were performed in Clustal Omega[51] and displayed using the online server of Espript 3.0[52].

In vitro DNA cleavage assay

Linear substrate DNA used in the cleavage experiments was generated by PCR using pBR322-based vectors as template and purified by agarose gel electrophoresis. Wild-type or mutant *Hz*Transib (300 nM final concentration), substrate DNA (final concentration 30 nM) were incubated in reaction buffer (25 mM MOPS, pH7.0, 50 mM KCl, 2 mM DTT, 5 mM $MgCl_2$; 16 µl final reaction volume) at 30 °C for 1 h. Reactions were stopped by adding 1.25 µl 2.5% SDS, 5 µl proteinase K (200 µg ml^{-1}) and 2 µl 0.5 M EDTA followed by incubation at 55 °C for 3 h. Samples were briefly centrifuged and the supernatant mixed with 6 µl 5× high-density Tris-borate-EDTA (TBE) sample buffer (ThermoFisher Scientific) and loaded on a non-denaturing 1× TBE-buffered polyacrylamide gel (Bio-Rad or ThermoFisher Scientific). After 35 min electrophoresis at 160 V, gels were stained with SYBR gold (ThermoFisher Scientific) in 1× TBE buffer for 1 h and imaged using a PharosFX Plus (Bio-Rad).

In vitro transposition assay

Linear donor DNA with tetracycline-resistant gene was amplified by PCR using the pBR322-based vector as template and purified by agarose gel electrophoresis. 0.05 pmol donor DNA and 0.1 pmol pECFP-1 target plasmid were mixed with 150 ng wild-type *Hz*Transib protein in reaction buffer (25 mM MOPS, pH7.0, 50 mM KCl, 2 mM DTT, 5 mM $MgCl_2$) and incubated at 30 °C for 1 h. After protease K digestion, DNA was ethanol-precipitated. 200 ng of DNA was transformed into electrocompetent MC1061 bacterial cells that were spread onto plates containing kanamycin or kanamycin + tetracycline + streptomycin (KTS)[13]. Plasmids from 54 colonies from KTS plates were sequenced to determine the integration location on the plasmid and target site duplication (TSD) sequence. Sequence logo representing nucleotide frequencies of *Hz*Transib TSD were generated and visualized with kpLogo web server[53].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Atomic coordinates of six *Hz*Transib or *Hz*Transib–TIR DNA complex structures have been deposited in PDB under accession number 6PQN (*Hz*Transib apo), 6PQR (*Hz*Transib intact TIR PRC), 6PQU (*Hz*Transib nicked TIR PRC), 6PQX (*Hz*Transib TIR HFC), 6PQY (*Hz*Transib TIR TEC) and 6PR5 (*Hz*Transib TIR STC). Five cryo-EM density maps of *Hz*Transib

complexed with different TIR DNA have been deposited in the Electron Microscopy Data Bank under accession numbers EMD-20452, EMD-20453, EMD-20455, EMD-20456 and EMD-20457, respectively.

37. Kabsch, W. Xds. *Acta Crystallogr. D* **66**, 125–132 (2010).
38. Sheldrick, G. M. A short history of SHELX. *Acta Crystallogr. A* **64**, 112–122 (2008).
39. Terwilliger, T. C. et al. Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. *Acta Crystallogr. D* **65**, 582–601 (2009).
40. Terwilliger, T. C. et al. Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr. D* **64**, 61–69 (2008).
41. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
42. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
43. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
44. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
45. Rohou, A. & Grigorieff, N. CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
46. Zivanov, J. et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7**, e42166 (2018).
47. Bai, X. C., Rajendra, E., Yang, G., Shi, Y. & Scheres, S. H. Sampling the conformational space of the catalytic subunit of human γ-secretase. *eLife* **4**, e11182 (2015).
48. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).
49. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
50. Goddard, T. D. et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).
51. Sievers, F. & Higgins, D. G. Clustal omega. *Curr. Protoc. Bioinformatics* **48**, 3.13.1–3.13.16 (2014).
52. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320–W324 (2014).
53. Wu, X. & Bartel, D. P. kpLogo: positional *k*-mer analysis reveals hidden specificity in biological sequences. *Nucleic Acids Res.* **45**, W534–W538 (2017).

Acknowledgements We thank W. Eliason for assistance with size-exclusion chromatography–multiple angle light scattering; K. Zhou for assistance in freezing the cryo-EM grids of *Hz*Transib-intact TIR complex; S. Wu for help with cryo-EM data collection at Yale West Campus; the staff of the Advanced Photon Source beamlines 24-ID-C and 24-ID-E for technical assistance with X-ray crystallography data collection; N. Craig for critical reading and many helpful comments on the manuscript. We are grateful for the advice, mentoring and support from T. Steitz during the early phases of this work. This work was supported by NIH grant R01 AI137079 (D.G.S.), Yale University School of Medicine James Hudson Brown-Alexander Brown Coxe Postdoctoral Fellowship (C.L.) and NVIDIA GPU Grant Program (C.L. and Y.Y.).

Author contributions C.L. and D.G.S. conceived the project and designed the experiments. C.L. performed cloning, protein expression, purification, complex reconstitution, sample screening using negative-stain electron microscopy, cryo-EM grids preparation and functional assays. C.L. and Y.Y. carried out protein crystallization, crystal structure determination, cryo-EM data collection and processing, atomic model building and refinement. C.L. and D.G.S. analysed data and wrote the manuscript with input from Y.Y.

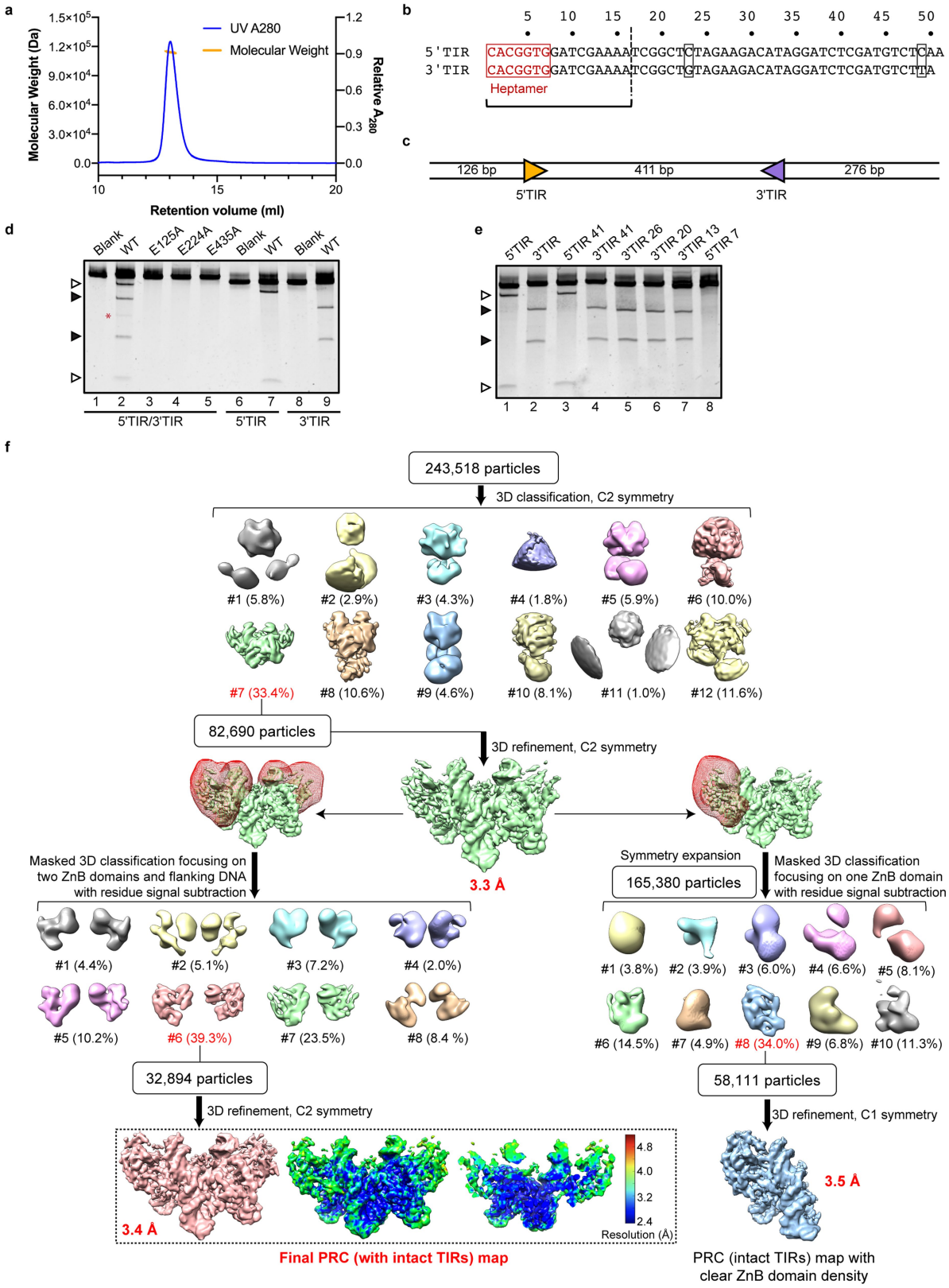
Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1753-7.

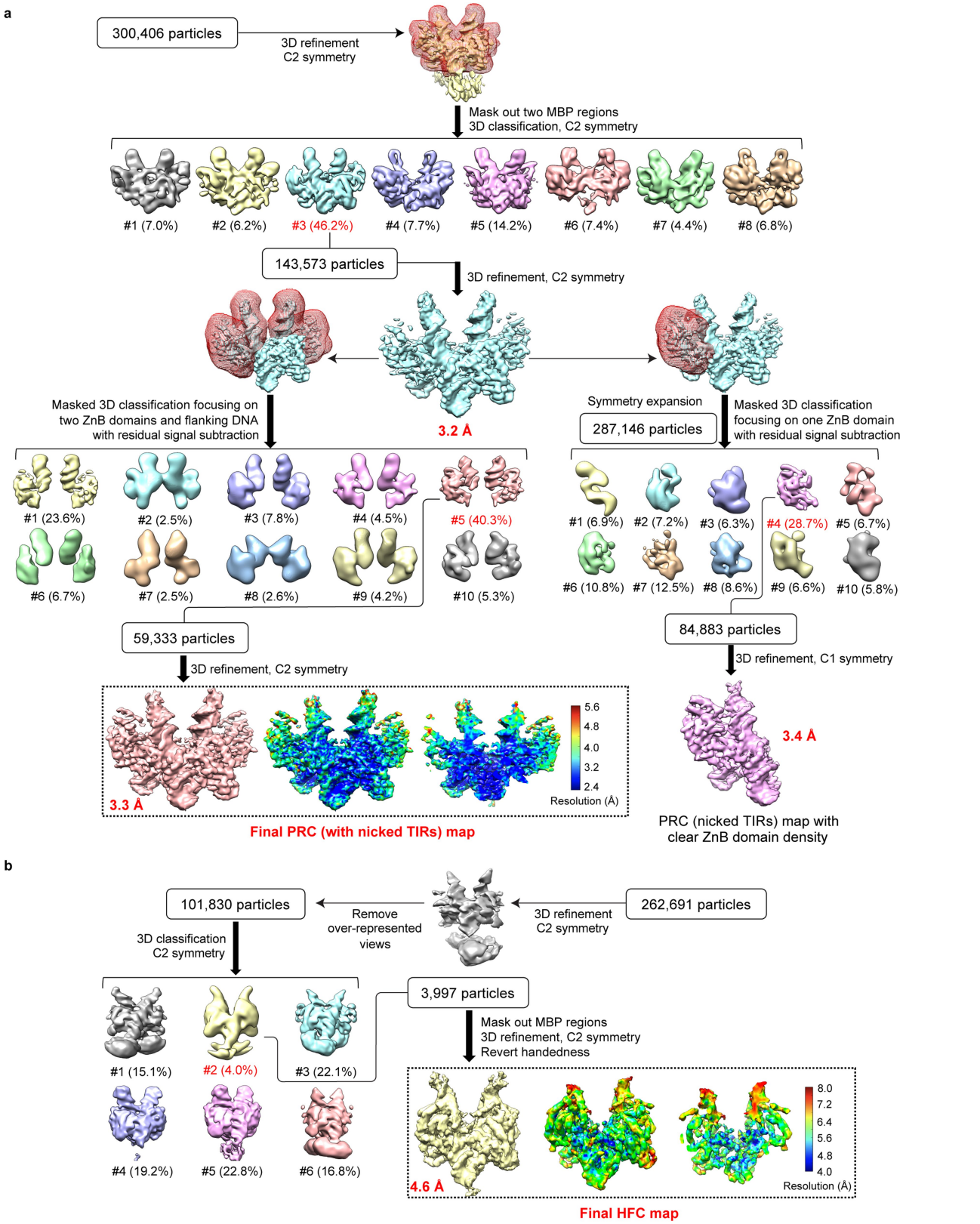
Correspondence and requests for materials should be addressed to D.G.S.

Peer review information *Nature* thanks Orsolya Barabas, Thomas Boehm, Ronald Chalmers and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

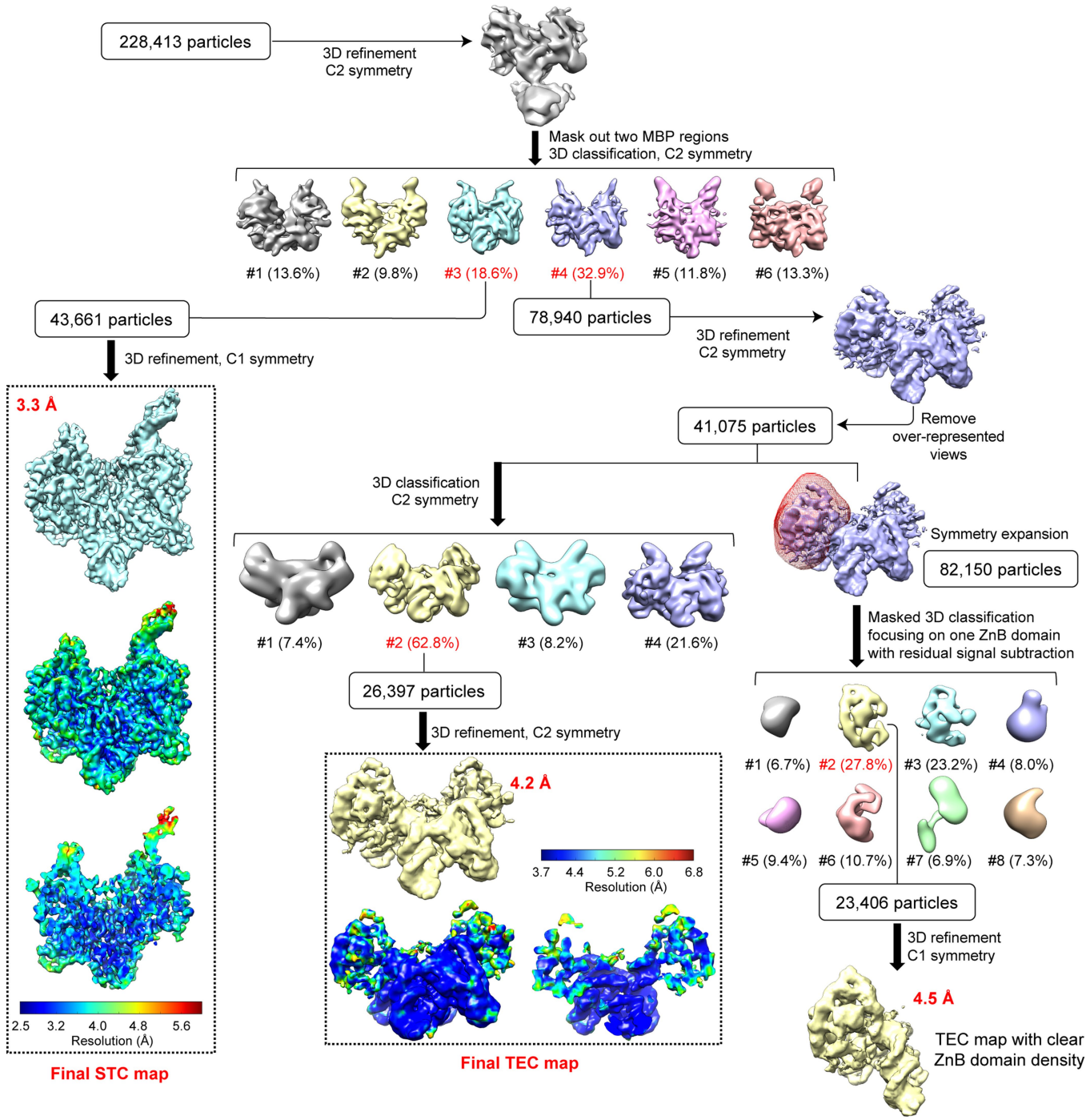


Extended Data Fig. 1 | See next page for caption.

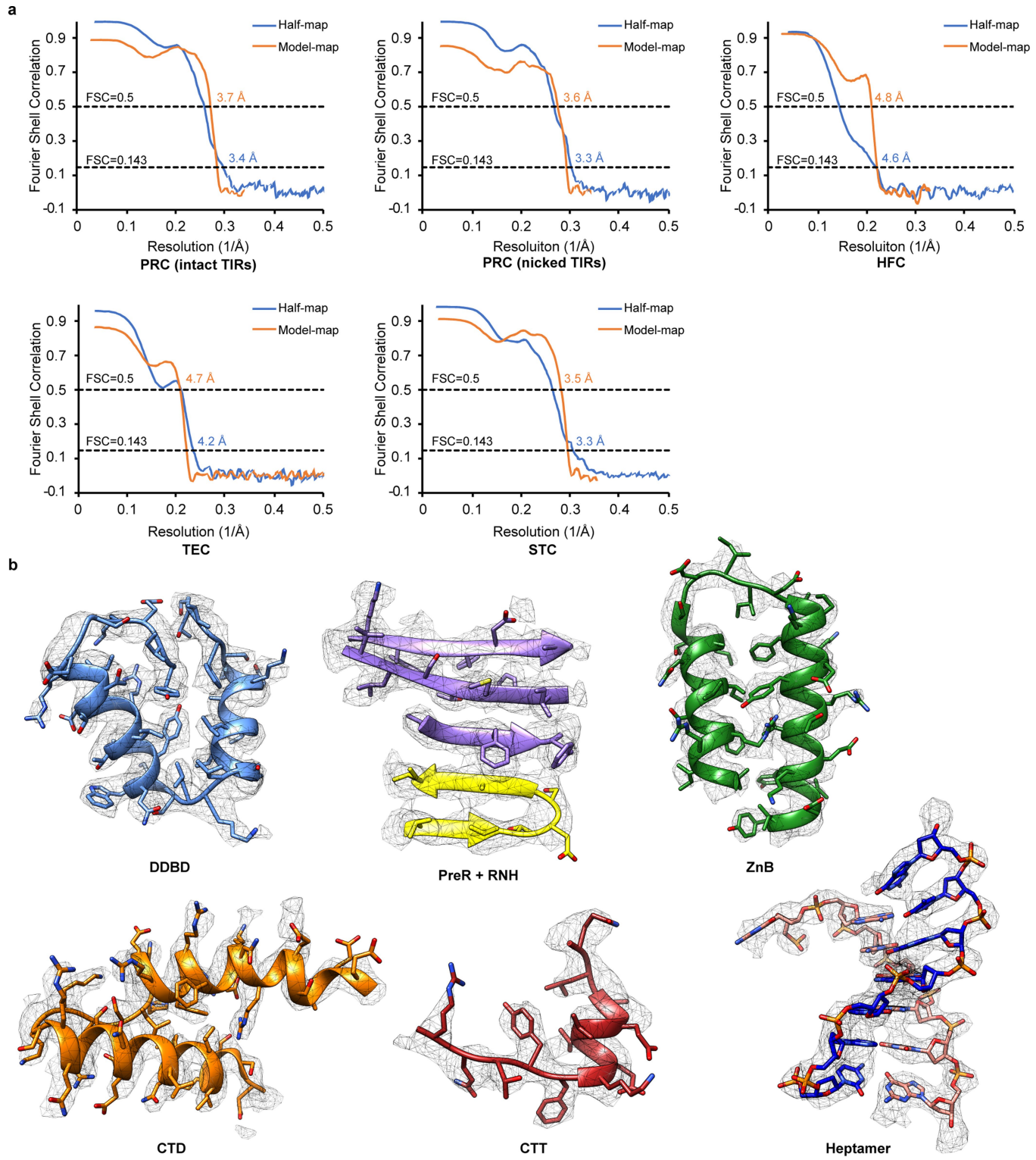
Extended Data Fig. 1 | Biochemical characterization of *H. zea* Transib transposase and single-particle cryo-EM analysis of Transib in complex with intact TIR substrates. a, Size-exclusion chromatography–multiple angle light scattering analysis of purified Transib protein, indicating that it forms a dimer in solution. Size-exclusion chromatography was repeated three times and similar profiles were obtained. The multiple angle light scattering experiment was not repeated. **b**, Numbering and sequence of endogenous left end (5′TIR) and right end (3′TIR) of the Transib transposon with nucleotide differences in black boxes. The first 16 bp of the TIR sequence are the same as the 16-bp transposon end of the TIR substrates used in structure determination. **c**, Schematic of the TIR substrate DNA used in the in vitro DNA-cleavage assay. 5′TIR and 3′TIR are shown as yellow and purple triangles, respectively. **d**, Cleavage of DNA substrates bearing one or two TIRs by MBP-tagged wild-type or mutant Transib transposases, each with the N-terminal 16 amino acids removed. The experiment was repeated three times and similar results were obtained. For gel source data, see Supplementary Fig. 1. **e**, Cleavage of DNA substrates bearing either full-length (lanes 1 and 2) or truncated (lanes 3–8) 5′TIR or 3′TIR, with site of truncation indicated in the substrate name. The experiment was repeated three times and similar results were obtained. Open and closed arrowheads indicate single 5′TIR and single 3′TIR cleavage products, respectively. Red asterisk marks the double cleavage band. The DNA cleavage products were resolved in 5% TBE polyacrylamide gels and stained with SYBR Gold. **f**, Flow chart of cryo-EM structure determination of Transib in complex with intact TIR substrates. After the first round of 3D classification, 3D auto-refinement using all of the particles in the best class generated a 3.3 Å map. Further 3D classifications focusing on either two ZnB domains plus flanking DNA regions or on one ZnB domain with symmetry expansion were used to obtain the final 3.4 Å map or a 3.5 Å map with clear ZnB domain density. All three maps were used as cross-references for model building. The final map and accompanying local resolution illustrations are enclosed in the dashed black box.



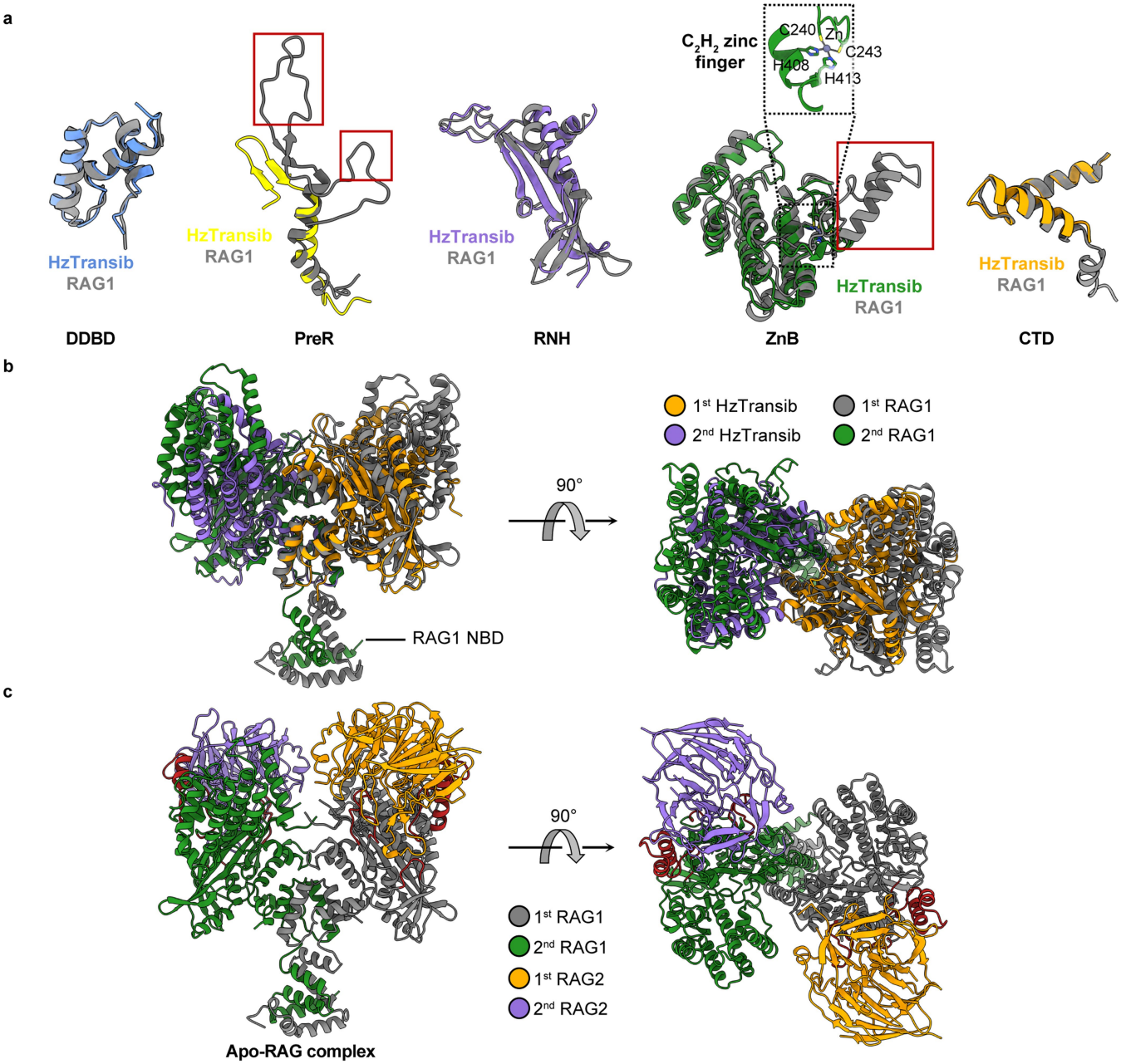
Extended Data Fig. 2 | Single-particle cryo-EM analysis of Transib in complex with nicked TIR substrates. a, Flow chart of cryo-EM image processing for Transib PRC with nicked TIR substrates. After the first round of 3D classification with two MBP regions masked, 3D auto-refinement using all of the particles in the best class generated a 3.2 Å map. Further 3D classifications focusing on either two ZnB domains plus flanking DNA regions or on one ZnB domain with symmetry expansion were used to obtain the final 3.3 Å map or a 3.4 Å map with clear ZnB domain density. All three maps were used as cross-references for model building. **b**, Flow chart of cryo-EM image processing for Transib HFC with nicked TIR substrates. After initial 3D auto-refinement, particles in the over-represented 2D classes were manually adjusted to alleviate the preferred particle orientation problem. Subsequent 3D classification and auto-refinement yielded a 4.6 Å map with even angular distribution. The final maps and accompanying local resolution illustrations are enclosed in the dashed black box.



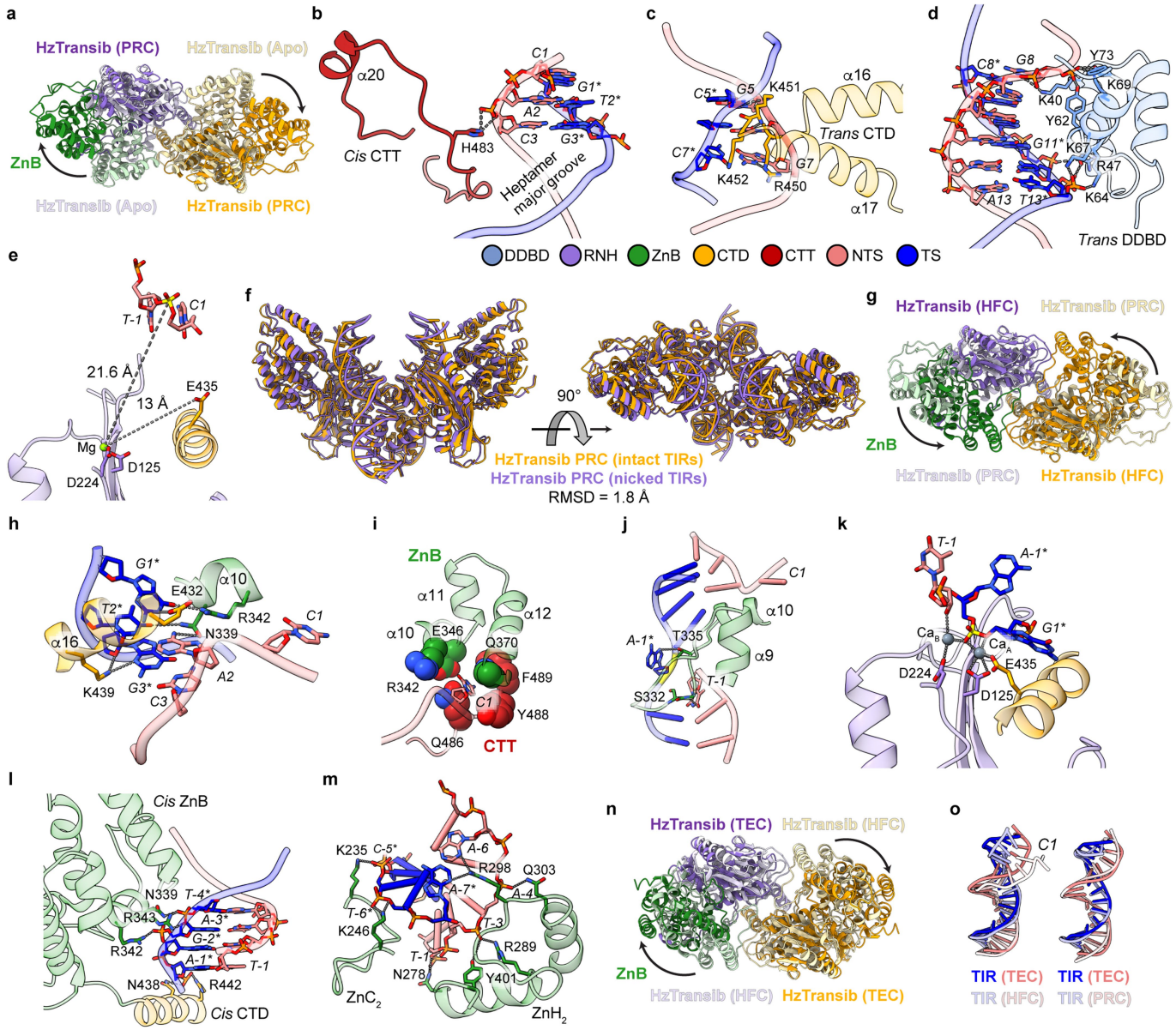
Extended Data Fig. 3 | Single-particle cryo-EM analysis of Transib in complex with TIR substrates in reaction conditions that support catalysis. Flow chart for Transib TEC and STC map reconstructions from Transib–intact TIR DNA complex prepared at 30 °C in the presence of Mg^{2+}. Different subsets of particle images were selected from different classification schemes to produce three refined cryo-EM maps: final STC map at 3.3 Å, final TEC map at 4.2 Å and a 4.5 Å map encompassing one Transib and TIR protomer in TEC with clear ZnB domain density. The two TEC maps were used as cross-references for model building. The final STC and TEC maps and accompanying local resolution illustrations are enclosed in the dashed black box.



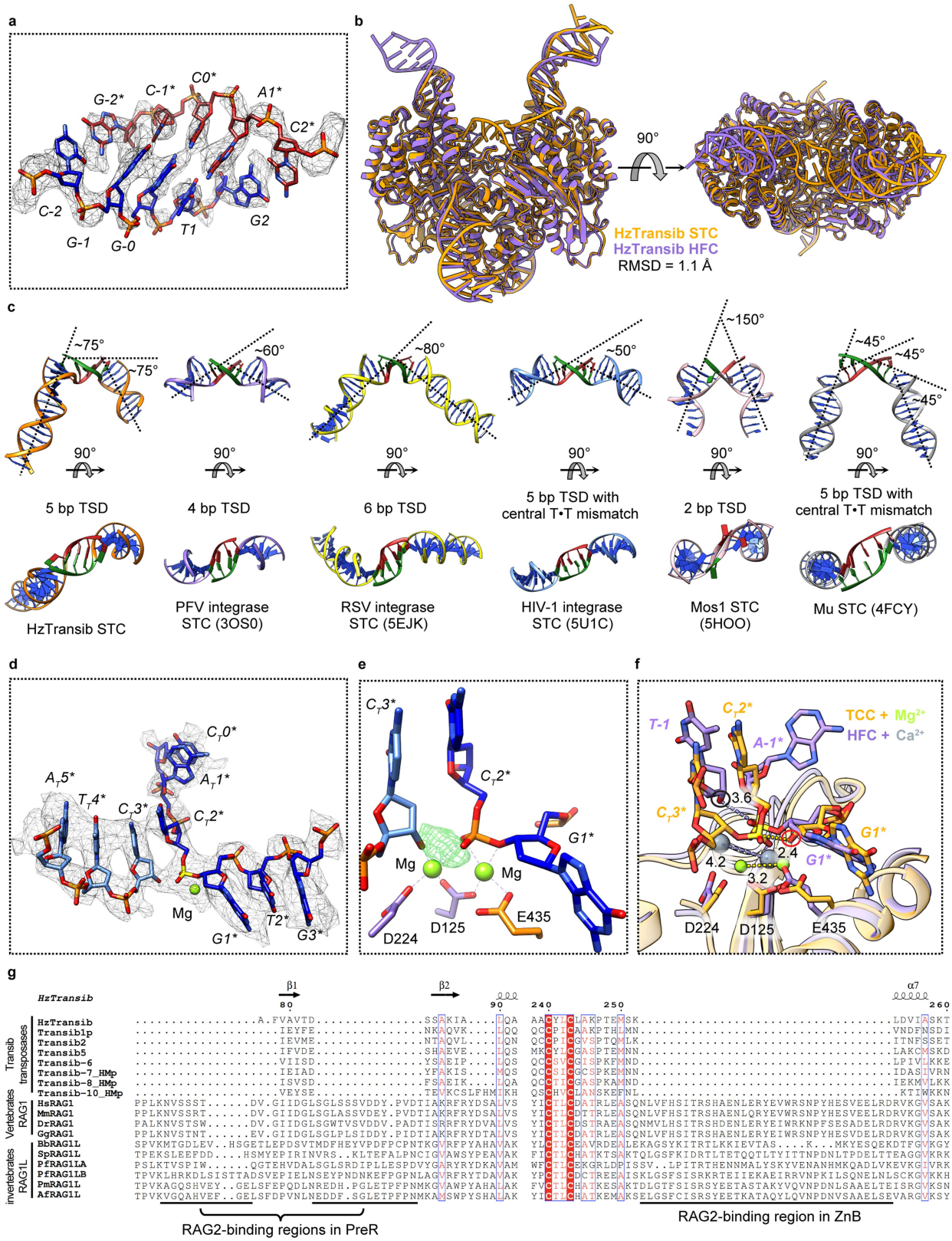
Extended Data Fig. 4 | Validation of cryo-EM structural models. a, Half-map FSC and model-map FSC curves of five cryo-EM maps from this study were generated from MolProbity. Gold-standard FSC curves between the two half maps with indicated resolution at FSC = 0.143 are in blue. FSC curves between the atomic model and the final map with indicated resolution at FSC = 0.5 are in orange. **b**, Cryo-EM densities superimposed on the atomic model for representative regions of Transib and TIR complexes.



Extended Data Fig. 5 | Structural comparison of Transib with RAG1. **a**, Superimposition of individual domains from Transib and RAG1 structures. Because the ZnC_2 portion of the ZnB domain is missing from the Transib apo structure, the ZnB domain from Transib STC was used for structural superimposition. Three structural motifs in RAG1 that are responsible for RAG2 interactions are highlighted in red boxes. **b**, The front and top views of Transib and RAG1 dimer superimposed by their DDBD domains. **c**, Front and top view of the apo RAG1-RAG2 heterotetramer structure (PDB 4WWX)[8].

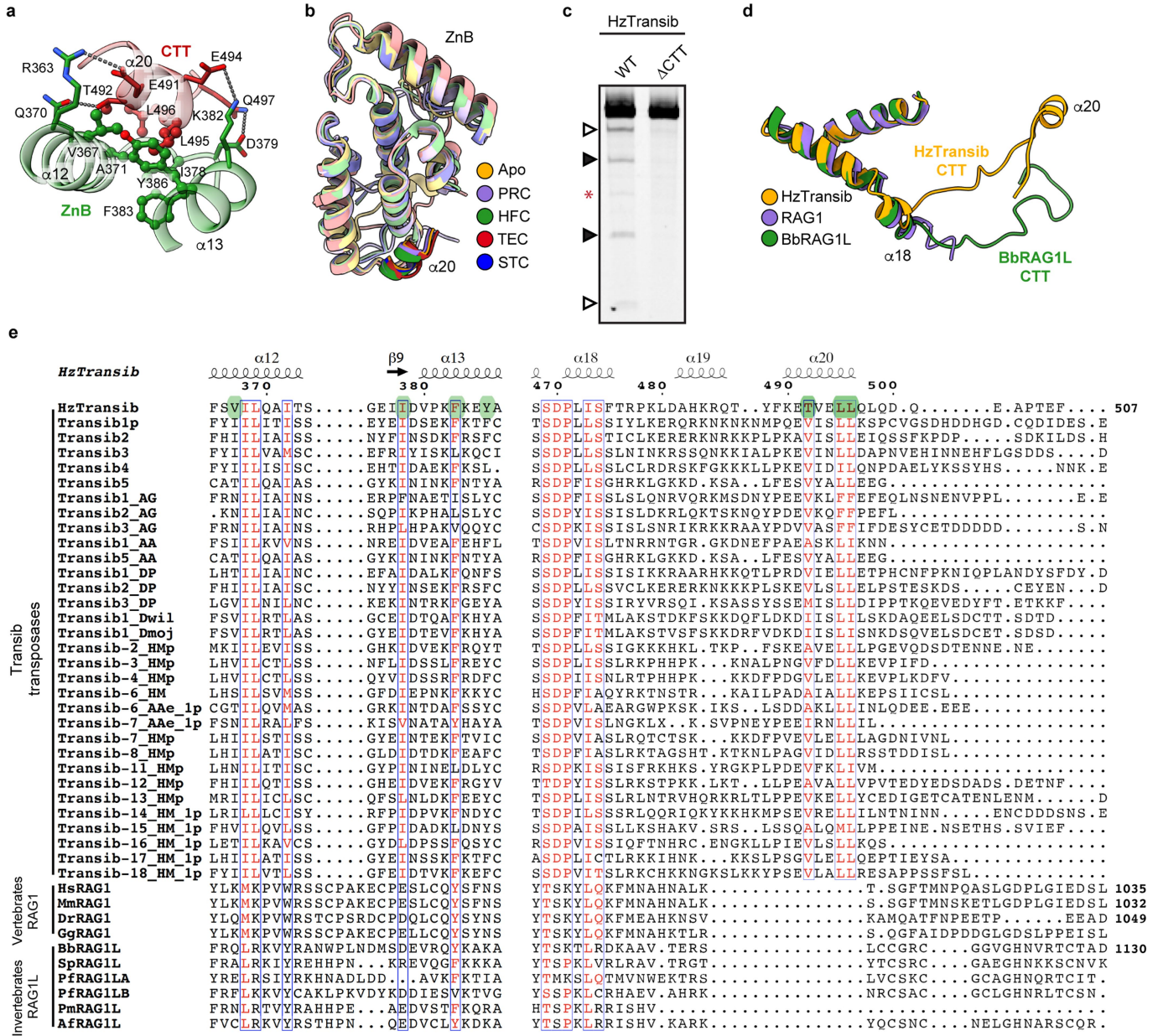


Extended Data Fig. 6 | TIR recognition in Transib PRC, HFC and TEC. **a**, Superimposition of Transib dimer in PRC (dark colours) and apo (pale colours) structures by their DDBD illustrates the large conformational changes of ZnB domains (green in one subunit). **b**–**e**, TIR recognition in Transib PRC. **b**, Interactions between Transib CTT and the heptamer. Hydrogen bonds are shown as grey dotted lines. Labels for nucleotide residues are italic. **c**, Interactions between Transib and last three base pairs of heptamer. **d**, Interactions between Transib and transposon end DNA downstream of heptamer. **e**, Active site of Transib PRC structure. Distances between Mg^{2+} ion and scissile phosphate or E435 are indicated. **f**, The front and top views of two Transib PRC structures (incubated with either intact or nicked TIRs at 4 °C) superimposed by their DDBD domains. The Transib nicked PRC complex is referred to as a PRC because of its strong structural resemblance to the intact DNA PRC. Depending on reaction conditions (temperature and divalent cation; see Methods), the nicked TIR substrate can be incorporated into either a nicked PRC or the HFC. **g**, Superimposition of Transib dimer in HFC and PRC structures by their DDBD shows the inward movements of ZnB domains and dimer closure. **h**–**k**, TIR recognition in Transib HFC. **h**, Interactions between Transib and the first three base pairs of heptamer. **i**, The first nucleotide of the heptamer (C1) is flipped out and buried in a pocket. **j**, Interactions between Transib α9–α10 loop and TIR at heptamer-flanking DNA junction. **k**, Active site of Transib HFC structure. **l**, Interactions between Transib and TIR flanking DNA in PRC. **m**, Interactions between Transib ZnB domain and TIR flanking DNA in HFC. **n**, Superimposition of Transib dimer in TEC and HFC structures by their DDBD shows the outward movements of ZnB domains. **o**, Comparison of transposon end DNA in TEC to that in HFC or in PRC. Mg^{2+} and Ca^{2+} ions are green and slate grey, respectively; other structure elements are coloured as in Fig. 2b. Scissile phosphate in each structure is highlighted in yellow.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Validation and analysis of Transib STC structure. **a**, Superimposition of 5-bp TSD region with the cryo-EM map contoured at 5.5σ. **b**, Front and top views of Transib STC structure superimposed on the Transib HFC structure. **c**, Comparison of target DNA from *H. zea* Transib, retrovirus integrases, Mos1 transposase and Mu transposase STC structures. Target site DNAs are shown as green and red. The approximate degree of bending in each target DNA is indicated. *H. zea* Transib is the only DDE/D-family transposase–integrase for which a STC structure has been reported that lacks a bend or base-unpairing at the centre of the target site DNA. Instead, Transib strongly bends target DNA near both edges of the target site DNA (between position −2 and −1 and positions 1 and 2), leading to a total 150° directional change of target DNA. Target DNAs in retroviral integrase STC structures exhibit relatively mild bends with one backbone kink at the centre of target site DNA, regardless of its length (ranging from 4 bp in PFV integrase to 6 bp in RSV integrase). The sharp bending (about 150°) at the centre of the Mos1 target DNA is achieved by flipping of the adenines in the TA target site. The target DNA in Mu STC exhibits a more continuous bending pattern through the 5-bp target site DNA, with one bend before the target site (between position −3 and −2), one at the centre and one immediately after the target site DNA (between position 2 and 3). The central bend is facilitated by the T–T mismatch in the target site. **d**, Transposon end–target DNA junction region of the Transib STC model superimposed on the cryo-EM map contoured at 5.5σ. Nucleotide residues in target DNA are labelled with a subscript T. **e**, Difference density between the Transib STC cryo-EM map and the model, showing the uncleaved target DNA phosphodiester bond in a portion of the particles used for cryo-EM map reconstruction. The difference map was contoured at 6σ. **f**, Superimposition of Transib TCC (protein in orange and metal ions in green) active site with Transib HFC active site (protein in purple and metal ions in grey). Distances are expressed in Å. Attacking oxygen atoms in HFC and TCC are highlighted in black and red circles, respectively. In TCC, the phosphorus is 2.4 Å from the attacking oxygen and the two metal ions are 3.2 Å apart. These distances are 3.6 Å and 4.2 Å in HFC. **g**, Sequence alignment of Transib transposases, vertebrate RAG1 and deuterostome invertebrate RAG1L proteins, showing the regions corresponding to three RAG2-binding interfaces in RAG1. Residue numbers are for *H. zea* Transib. Hs, *Homo sapiens* (human); Mm, *Mus musculus* (mouse); Dr, *Danio rerio* (zebrafish); Gg, *Gallus gallus* (chicken); Bb, *Branchiostoma belcheri* (amphioxus); Sp, *Strongylocentrotus purpuratus* (purple sea urchin); Pf, *Ptychodera flava* (acorn worm); Pm, *Petromyzon marinus* (sea lamprey) and Af, *Asterias forbesi* (sea star).



Extended Data Fig. 8 | Structural insights into the function and evolution of *H. zea* Transib CTT. a, Interactions between Transib CTT α20 and ZnB domain α12–α13. Residues in CTT and ZnB are coloured red and green, respectively. Residues involved in hydrophobic interactions are shown in ball-and-stick representation. **b**, Superimposition of ZnB domain (pale colours) together with CTT α20 (dark colours) from the structures representing five steps in transposition. **c**, Cleavage of DNA substrates bearing a 5′TIR–3′TIR pair by MBP-tagged wild-type or CTT truncated mutant Transib transposases, each with N-terminal 16 amino acids removed. The DNA cleavage products were resolved on a 6% TBE polyacrylamide gel and stained with SYBR Gold. Open and closed arrowheads indicate single 5′TIR and single 3′TIR cleavage products, respectively. The red asterisk marks the double-cleavage band. The experiment was repeated at least three times independently and similar results were obtained. For gel source data, see Supplementary Fig. 1. **d**, Superimposition of Transib, RAG1 and *Bb*RAG1L structures by the first two helices of their CTDs. Transib and *Bb*RAG1L CTT extend from the structurally conserved CTD and point in different directions. **e**, Sequence alignment of *H. zea* Transib CTT with vertebrate RAG1 CTT and deuterostome invertebrate RAG1L CTT showing highly divergent sequences among the three groups. Residues mediating the hydrophobic interactions between ZnB α12–α13 and CTT α20 are highlighted in green. Residue numbers and secondary structure elements at the top of the sequence alignment are for *H. zea* Transib. The residue number for the final amino acid in the sequence alignment is indicated for selected sequences.

Extended Data Table 1 | Statistics of crystal data collection, phasing and refinement

	HzTransib apo native (PDB 6PQN)	HzTransib apo Br derivative	HzTransib apo I derivative	HzTransib apo Os derivative	HzTransib apo Pt derivative	HzTransib apo Hg derivative
Data collection						
Space group	$P6_122$	$P6_122$	$P6_122$	$P6_122$	$P6_122$	$P6_122$
Cell dimensions						
a, *b*, *c* (Å)	160.292, 160.292, 235.858	159.315, 159.315, 236.805	159.797, 159.797, 238.293	160.238, 160.238, 235.817	160.912, 160.912, 236.643	160.353, 160.353, 236.264
α, β, γ (°)	90, 90, 120	90, 90, 120	90, 90, 120	90, 90, 120	90, 90, 120	90, 90, 120
Resolution (Å)	200.0–3.01 (3.19–3.01)*	200.0–3.84 (4.30–3.84)	200.0–3.81 (4.17–3.81)	200.0–3.18 (3.35–3.18)	200.0–3.96 (4.42–3.96)	200.0–3.14 (3.31–3.14)
R_{sym} or R_{merge}	0.086 (1.829)	0.280 (3.748)	0.196 (3.157)	0.135 (3.403)	0.141 (2.482)	0.134 (3.983)
$I / \sigma I$	14.47 (0.75)	21.46 (0.9)	19.87 (1.2)	21.18 (0.9)	22.45 (1.1)	21.4 (0.9)
Completeness (%)	99.3 (99.3)	99.7 (99.1)	100.0 (99.8)	100.0 (100.0)	99.7 (99.1)	99.8 (99.0)
Redundancy	7.7 (7.6)	12.7 (12.7)	19.2 (19.9)	19.3 (19.1)	12.7 (13.3)	19.5 (20.0)
Refinement						
Resolution (Å)	80.15–3.01 (3.117–3.01)					
No. reflections	35894 (3455)					
R_{work} / R_{free}	0.220/0.277					
No. atoms						
Protein	7288					
Ligand/ion	83					
Water	28					
B-factors						
Protein	147.85					
Ligand/ion	183.99					
Water	108.59					
R.m.s deviations						
Bond lengths (Å)	0.004					
Bond angles (°)	0.77					

*One crystal was used for each dataset; values in parentheses are for the highest-resolution shell.

Extended Data Table 2 | Statistics of cryo-EM data collection, refinement and validation

	PRC (intact TIR) (EMD-20452) (PDB 6PQR)	PRC (nicked TIR) (EMD-20453) (PDB 6PQU)	HFC (EMD-20455) (PDB 6PQX)	TEC (EMD-20456) (PDB 6PQY)	STC (EMD-20457) (PDB 6PR5)
Data collection and processing					
Magnification	130,000	130,000	130,000	130,000	130,000
Voltage (kV)	300	300	300	300	300
Electron exposure (e^{-}/Å^2)	50.8	52.2	52.2	54.4	54.4
Defocus range (μm)	-1.5 – -2.5	-1.5 – -2.5	-1.5 – -2.5	-1.5 – -2.5	-1.5 – -2.5
Pixel size (Å)	1.05	1.05	1.05	1.05	1.05
Symmetry imposed	C2	C2	C2	C2	C1
Initial particle images (no.)	243,518	300,406	262,691	228,413	228,413
Final particle images (no.)	32,984	59,333	3,997	26,397	43,661
Map resolution (Å)	3.4	3.3	4.6	4.2	3.3
FSC threshold	0.143	0.143	0.143	0.143	0.143
Map resolution range (Å)	2.4 – 5.2	2.4 – 5.6	4.0 – 8.0	3.7 – 6.8	2.5 – 5.6
Refinement					
Initial model used (PDB code)	6PQN	6PQN	6PQN	6PQN	6PQN
Model resolution (Å)	3.7	3.6	4.8	4.7	3.5
FSC threshold	0.5	0.5	0.5	0.5	0.5
Model resolution range (Å)	2.4 – 5.2	2.4 – 5.6	4.0 – 8.0	3.7 – 6.8	2.5 – 5.6
Map sharpening *B* factor (Å^2)	-90	-90	-126	-120	-90
Model composition					
Non-hydrogen atoms	9266	10004	10116	8690	10194
Protein residues	936	936	960	920	960
Nucleotides	88	124	120	64	124
Ligands	6	6	4	0	6
B factors (Å^2)					
Protein	105.51	74.58	89.52	182.34	59.68
Nucleic acid	104.92	131.57	160.90	158.19	91.35
Ligand	109.58	93.13	81.45	–	69.54
R.m.s. deviations					
Bond lengths (Å)	0.009	0.008	0.006	0.007	0.008
Bond angles (°)	0.830	0.890	0.984	0.980	0.768
Validation					
MolProbity score	1.91	2.04	2.39	2.35	1.64
Clashscore	16.01	23.12	16.06	24.55	13.64
Poor rotamers (%)	0.24	1.21	2.35	0.49	0.71
Ramachandran plot					
Favored (%)	96.74	97.39	94.14	92.51	98.33
Allowed (%)	3.26	2.39	5.23	6.83	1.36
Disallowed (%)	0	0.22	0.63	0.66	0.31

nature research

Corresponding author(s): David G. Schatz

Last updated by author(s): Sep 19, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	SerialEM version 3.6
Data analysis	RELION-3.0; MotionCor2; CTFFIND-4.1.10; Gautomatch 0.56; Phenix-1.15; Coot-0.8.8; ResMap-1.1.4; UCSF Chimera 1.13; UCSF ChimeraX 0.9; PyMol 2.0; XDS version March 2018; SHELXD-2013; AutoSol, AutoBuild and MolProbity contained in Phenix-1.15; kpLogo web server

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Atomic coordinates of six HzTransib or HzTransib-TIR DNA complex structures have been deposited in PDB under accession number 6PQN (HzTransib Apo), 6PQR (HzTransib-intact TIR PRC), 6PQU (HzTransib-nicked TIR PRC), 6PQX (HzTransib-TIR HFC), 6PQY (HzTransib-TIR TEC) and 6PR5 (HzTransib-TIR STC). Five cryo-EM density maps of HzTransib complexed with different TIR DNA have been deposited in the Electron Microscopy Data Bank under accession number EMD-20452, EMD-20453, EMD-20455, EMD-20456, EMD-20457, respectively. The atomic coordinates and cryo-EM density maps will be publicly available prior to publication.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. Sufficient cryo-EM data were collected to achieve adequate map resolutions for model building. The sample size was based on previous studies.
Data exclusions	No data were excluded from analyses.
Replication	All biochemical assays were repeated three times. Each replicate was an independent experiment and did not represent re-assay of the same material. All attempts at replication were successful.
Randomization	Randomization was not relevant to our study because our study did not involve the allocation of samples/organisms/participants into experimental groups.
Blinding	Investigators were not blinded to group allocation because group allocation was not involved in our study. Investigators were not blinded during data collection because the data being collected were quantitative in nature (gels or numbers of colonies on a plate) and were not prone to subjective interpretation.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Sf9 cells in Sf-900 III serum-free media were obtained from Thermofisher Scientific (catalog number 12659017)
Authentication	Cell lines used were not authenticated.
Mycoplasma contamination	Cell lines were not tested for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in this study.

Article

Architecture of autoinhibited and active BRAF–MEK1–14-3-3 complexes

https://doi.org/10.1038/s41586-019-1660-y

Received: 14 February 2019

Accepted: 26 September 2019

Published online: 3 October 2019

Eunyoung Park[1,2], Shaun Rawson[2], Kunhua Li[1,2], Byeong-Won Kim[1,2], Scott B. Ficarro[1,3,4], Gonzalo Gonzalez-Del Pino[1,2], Humayun Sharif[1,2], Jarrod A. Marto[1,3,4], Hyesung Jeon[1,2*] & Michael J. Eck[1,2*]

RAF family kinases are RAS-activated switches that initiate signalling through the MAP kinase cascade to control cellular proliferation, differentiation and survival[1–3]. RAF activity is tightly regulated and inappropriate activation is a frequent cause of cancer[4–6]; however, the structural basis for RAF regulation is poorly understood at present. Here we use cryo-electron microscopy to determine autoinhibited and active-state structures of full-length BRAF in complexes with MEK1 and a 14-3-3 dimer. The reconstruction reveals an inactive BRAF–MEK1 complex restrained in a cradle formed by the 14-3-3 dimer, which binds the phosphorylated S365 and S729 sites that flank the BRAF kinase domain. The BRAF cysteine-rich domain occupies a central position that stabilizes this assembly, but the adjacent RAS-binding domain is poorly ordered and peripheral. The 14-3-3 cradle maintains autoinhibition by sequestering the membrane-binding cysteine-rich domain and blocking dimerization of the BRAF kinase domain. In the active state, these inhibitory interactions are released and a single 14-3-3 dimer rearranges to bridge the C-terminal pS729 binding sites of two BRAFs, which drives the formation of an active, back-to-back BRAF dimer. Our structural snapshots provide a foundation for understanding normal RAF regulation and its mutational disruption in cancer and developmental syndromes.

RAF activity is restrained by an intricate interplay that involves phosphorylation events, binding to 14-3-3 proteins, and intramolecular autoinhibitory interactions[1,3]. The mammalian RAF kinases ARAF, BRAF and CRAF share three conserved regions (CR1, CR2 and CR3; Fig. 1a). The N-terminal CR1 region contains the RAS-binding domain (RBD) and the cysteine-rich domain (CRD), whereas the C-terminal CR3 region contains the serine/threonine kinase domain and a motif that, when phosphorylated, serves as a binding site for 14-3-3 proteins. The intervening CR2 region consists of a second 14-3-3 recognition site (Fig. 1a). 14-3-3s are dimeric proteins that bind specific serine- or threonine-phosphorylated motifs in diverse signalling proteins[7]. In the absence of activating interactions with RAS, RAF proteins are thought to be maintained in an autoinhibited state that involves intramolecular interaction of the CR1 region with the kinase domain, and the binding of 14-3-3 proteins to the phosphorylated CR2 (pS365 in BRAF) and C-terminal (pS729 in BRAF) 14-3-3 binding sites[1,3]. RAF is recruited to the plasma membrane and activated in a process that involves the binding of GTP-bound RAS to its RBD domain. The adjacent CRD is also important for RAS-driven recruitment and activation[8,9]. Structurally, the CRD is a C1 domain, a small modular domain found in many lipid- or membrane-activated signalling proteins[10]. Normal RAF activation requires dimerization of the kinase domain[11], and active RAFs form both homo- and heterodimers[12,13]. MEK1 and MEK2 are the only known RAF substrates, and MEKs in turn selectively phosphorylate ERK1 and ERK2, the terminal kinases in the RAS/MAP kinase cascade. Recent work has revealed that BRAF is pre-associated with MEK in the quiescent state[14,15]. Previous structural studies of BRAF and other family members have been restricted to isolated domains or fragments of these proteins[16–18]. To aid in developing an integrated structural understanding of the normal regulation of RAFs and their pathological activation in cancer, we sought to prepare and structurally characterize intact BRAF in autoinhibited and active states in complexes with MEK1 and 14-3-3 proteins.

Overall structure of autoinhibited BRAF

We co-expressed full-length wild-type BRAF with full-length MEK1 in insect (Sf9) cells. To eliminate potential heterogeneity due to phosphorylation of the MEK1 activation loop, we used a variant of MEK1 in which these phosphorylation sites were mutated to alanine (S218A/S222A) in our structural studies. Affinity purification of the expressed proteins yielded well-defined complexes that also contained insect-cell-derived 14-3-3ε,ζ dimers (see Methods and Extended Data Fig. 1a). Co-expression of human 14-3-3 isoforms with BRAF and MEK1 did not fully displace the insect-cell 14-3-3s and led to increased heterogeneity (Extended Data Fig. 1b). We therefore exploited the binding of the abundant and highly conserved endogenous 14-3-3 proteins. This approach enabled us to isolate 'monomeric' complexes that contained a single chain of each of BRAF, MEK1 and the two 14-3-3 subunits (Extended

[1]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA. [2]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA. [3]Blais Proteomics Center, Dana-Farber Cancer Institute, Boston, MA, USA. [4]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. *e-mail: hjeon@crystal.harvard.edu; eck@crystal.harvard.edu

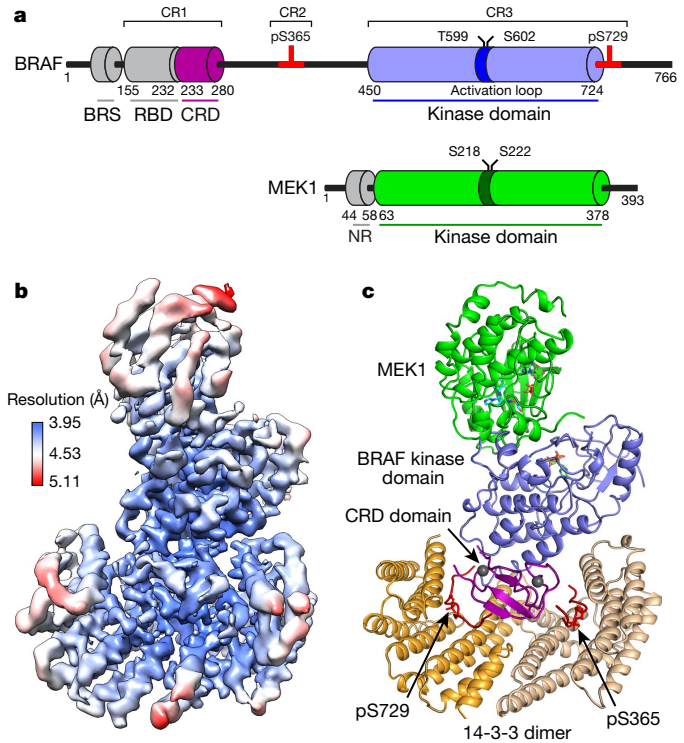


Fig. 1 | Structure of an autoinhibited BRAF–MEK1–14-3-3 complex. **a**, Schematic showing the domain organization of BRAF and MEK1. Key regulatory phosphorylation sites are indicated above the schematics and residue numbers for domain boundaries are shown below. BRS, BRAF-specific domain, which is unique to BRAF; NR, MEK negative regulatory region. **b**, Single-particle reconstruction cryo-EM density map derived from imaging of the full-length BRAF–MEK1–14-3-3 complex, coloured according to local resolution. **c**, Ribbon diagram showing the overall structure of the complex. BRAF and MEK1 domains are coloured as in **a**, and the two subunits of the 14-3-3 dimer are shown in orange and tan. Segments of BRAF containing the pS365 and pS729 regulatory sites bind opposite sides of the 14-3-3 dimer, and are shown in red. The CRD occupies a central location in the complex, and has contacts to both 14-3-3 subunits, both the pS365 and pS729 regulatory segments, and the BRAF kinase domain.

Data Fig. 1a). Consistent with an autoinhibited state, we found that both the S365 and S729 sites on BRAF were highly phosphorylated whereas those on the activation segment were not (Extended Data Fig. 1c). In complexes with wild-type MEK1, the MEK1 activation segment sites were also predominantly unphosphorylated (Extended Data Fig. 1c).

We prepared a 192-kDa BRAF–MEK1–14-3-3ε,ζ complex in the presence of the ATP analogue ATP-γ-S and the MEK inhibitor GDC-0623. Cryo-electron microscopy (cryo-EM) imaging of this complex revealed well-dispersed particles, with two-dimensional (2D) class averages showing obvious secondary structure features (Extended Data Fig. 1d, e). Approximately 8,400 micrograph movies afforded single-particle reconstructions of this complex at a nominal resolution of 4.1Å (Fig. 1b, Extended Data Fig. 1f), as detailed further in Methods and Extended Data Table 1.

The cryo-EM map revealed a compact structure, in which inactive BRAF is secured in a 14-3-3 'cradle' by extensive interactions with the 14-3-3 dimer (Fig. 1c). The 14-3-3 engages both cognate sites in BRAF: the phosphorylated CR2 site (pS365) is bound in the recognition groove on one side of the 14-3-3 dimer, whereas the C-terminal (pS729) motif is bound in the groove on the opposite side of the dimer. The BRAF CRD domain is particularly central to the overall architecture of the complex. It contacts both subunits of the 14-3-3 dimer, both the pS365 and the pS729 binding motifs, and the C-lobe of the BRAF kinase domain. The BRAF kinase domain is oriented such that its active site faces away from the 14-3-3 domain, enabling it to coordinate MEK1 in a 'face-to-face' orientation. Both the MEK1 and the BRAF kinase domains exhibit stereotypical inactive conformations, in which their regulatory αC-helices are displaced from their active positions. The N-terminal BRAF-specific domain and the RBD of BRAF are not clearly defined in the cryo-EM map. A reconstruction that was filtered to a resolution of 5 Å and contoured at a lower level revealed density adjacent to the CRD that corresponds to the RBD domain (Extended Data Fig. 1g), but it did not provide sufficient detail to enable the positioning of an RBD model.

The 14-3-3 dimer organizes inactive BRAF

Our BRAF complexes contain an approximately equimolar ratio of the ε and ζ isoforms (Extended Data Fig. 1a), and we expect that each side of the 14-3-3 dimer is a mixture of the two isoforms in our reconstructions. For simplicity and convenience, our model is constructed using the *Spodoptera frugiperda* 14-3-3ζ sequence for both subunits, but with residue numbering corresponding to the human 14-3-3ζ isoform. The 14-3-3 dimer interacts with every ordered domain of BRAF, and the interacting residues are highly conserved across all 14-3-3 isoforms (Extended Data Fig. 2a). The most N-terminal portion of BRAF that is well defined in the cryo-EM maps is the CRD domain. The CRD domain fold, which is approximately 50 residues in length, contains a small β-sheet and is stabilized by two zinc coordination sites (Fig. 1c). The domain binds in the centre of the 14-3-3 cradle, with contacts to both subunits of the dimer (Extended Data Fig. 2b). Notably, two loops of the CRD domain that are expected to mediate association of the domain with the membrane[17,19,20] (residues 239–245 and 253–260) make extensive contact with the 14-3-3 domain in the autoinhibited complex (Extended Data Fig. 3a). Previous mutagenesis studies[21] of the CRAF CRD have identified two residues in this region that are important for binding to 14-3-3. The corresponding residues in BRAF (R239 and T241) are indeed found at the interface with the 14-3-3 domain (Extended Data Fig. 3a).

The poorly conserved linker that connects the CRD domain and the CR2 region is not visible in our map, but the phosphorylated CR2 segment is well defined in the phosphopeptide recognition groove on one side of the 14-3-3 dimer (Extended Data Fig. 3b). The ordered CR2 segment extends from Q359 to I371, with pS365 roughly at its centre. Beyond I371, the linker that connects CR2 to the kinase domain is not visible. The BRAF kinase C-lobe contacts both 14-3-3 subunits but interacts most extensively with the pS365-binding subunit, packing against its α9 helix and α8–α9 loop (Extended Data Fig. 3c). This portion of the 14-3-3 domain also contacts H510 and adjacent residues in the N-terminal lobe of the kinase domain. We observe continuous density connecting the C terminus of the BRAF kinase domain with the pS729 14-3-3 binding motif, which occupies the recognition groove on the opposite side of the 14-3-3 dimer (Extended Data Fig. 3d, e). BRAF residues S732–A736 thread between the CRD and the 14-3-3 domain as they exit the recognition groove, and weak density corresponding to a few additional residues indicates that the BRAF C terminus passes across a hydrophobic surface on the CRD domain before it becomes substantially disordered. The interactions of the 14-3-3 domain with the CR2 segment in the present structure are similar to those observed in a crystal structure of human 14-3-3ζ in complex with a CRAF peptide[22] (Extended Data Fig. 3f).

The autoinhibited BRAF–MEK kinase module

The BRAF and MEK1 kinase domains bind with their active-site clefts juxtaposed, and both kinases exhibit inactive conformations (Fig. 2a). Density for ATP-γ-S is visible in the BRAF active-site cleft (Extended Data Fig. 4a); in the MEK active site we also observe density corresponding to a bound nucleotide, which is seemingly ADP (Extended Data Fig. 4b). The inactive conformation of MEK1 and the face-to-face kinase orientation seen here is similar to that previously observed for the isolated

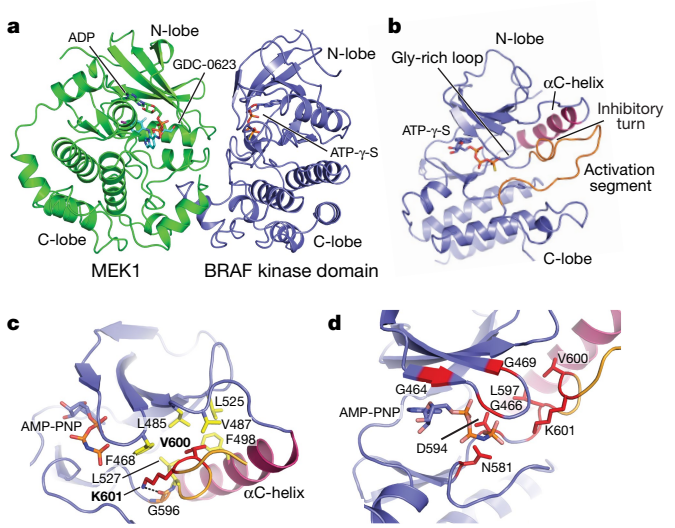


Fig. 2 | Conformation of the autoinhibited BRAF kinase domain and location of oncogenic mutations. a, BRAF coordinates MEK1 in a face-to-face orientation, with extensive contact between the kinase C-lobes. Both kinases adopt an inactive, αC-out conformation. **b**, Overall view of the autoinhibited BRAF kinase domain. The C-helix (magenta) is propped in an outward, inactive conformation by the inhibitory turn in the activation segment (orange). ATP-γ-S is bound in the active-site cleft. **c**, Detailed view of the structure and interactions of the inhibitory turn. Residue V600, the most common site of oncogenic mutations, is part of a cluster of hydrophobic residues (yellow) that stabilize this inactive conformation. K601 also stabilizes this configuration; it forms a hydrogen bond with G596 in the DFG motif and packs with F468 in the glycine-rich loop. **d**, Oncogenic mutations (red) cluster in the inhibitory turn or in residues that coordinate ATP. **a** and **b** are drawn from the cryo-EM structure; **c** and **d** from the crystal structure of the autoinhibited kinase domain complex.

kinase domains of BRAF and MEK1[14]. However, the BRAF kinase domain in the previous structure adopted an active but nucleotide-free conformation, as compared with the inactive, nucleotide-bound state in the present structure (Extended Data Fig. 4c, d). BRAF coordinates MEK1 through an extensive interface that involves primarily the C-lobes of both kinases, including the kinase activation segments, which interact in an antiparallel manner (Extended Data Fig. 4e, f).

In the BRAF kinase domain, the outward inactive position of the αC-helix is enforced by residues 598–602 in the activation segment, which form a helix-like turn that we refer to as the 'inhibitory turn' (Fig. 2b). The inhibitory turn packs together with hydrophobic residues in the glycine-rich loop, the C-helix and the β3 strand to stabilize the inactive state. This inhibitory arrangement resembles that observed in the inactive states of other kinases including CDK2, Src and EGFR. A superficially similar configuration has been observed in crystal structures of the BRAF kinase domain crystallized with sulfonamide-class inhibitors[23]; however, direct comparison with the present structure reveals marked differences in the activation segment and in the relative orientation of the N- and C-lobes of the kinase domain (Extended Data Fig. 4g, h). The inhibitor-bound conformation is approximately 15° more open as compared with the nucleotide-bound inactive state observed in the present structure. This difference prompted us to systematically examine the relative N- and C-lobe orientation in more than 50 BRAF kinase structures available in the Protein Data Bank. Notably, all previous BRAF kinase structures—none of which contain ATP or an ATP analogue—exhibit a markedly more open active-site cleft (owing to N-lobe rotations of 8–17°) as compared with the autoinhibited, nucleotide-bound structure described here (Extended Data Fig. 4i, j).

To obtain a higher-resolution view of the autoinhibited BRAF kinase domain, we co-expressed it with full-length MEK1(S218A/S222A) in insect cells and crystallized the purified complex in the presence of GDC-0623 and the ATP analogue AMP-PNP (Extended Data Fig. 5a). The resulting crystal structure, determined at a resolution of 2.6 Å, superimposes closely onto the corresponding portion of the autoinhibited BRAF–MEK1–14-3-3 cryo-EM structure (root mean square deviation, 0.56 Å; Extended Data Fig. 5b), and the regions of interest discussed above are highly similar. The crystal structure reveals in detail interactions that stabilize the inhibitory turn in the BRAF kinase domain (Fig. 2c), and interactions with the bound nucleotide (Extended Data Fig. 5c). Notably, oncogenic mutations in the BRAF kinase domain cluster in a small region that contains both the inhibitory turn and nucleotide-binding residues (Fig. 2d). Considering the extent of the interactions with the nucleotide and the unique N-lobe orientation observed in both the cryo-EM structure and the crystal structure, we propose that ATP binding is an essential feature of the autoinhibited state. We also observe a hydrogen bond between MEK1 residue E102 and the ribose group of the AMP-PNP bound to the BRAF kinase domain (Extended Data Fig. 5d).

The MEK1 portion of the crystal structure includes an N-terminal helix that is the site of rare activating mutations, both in cancer and in a 'RASopathy' known as cardio-cutaneo-facial syndrome—a genetic developmental disorder that stems from aberrant signalling in the MAP kinase pathway[24]. This α-helix packs across the back of the N-lobe (Extended Data Fig. 5b, e), apparently contributing to the stability of the inactive αC-out conformation of MEK, as seen in previous work[25]. For reasons that are not yet clear, this helix is not resolved in the cryo-EM map of the autoinhibited complex.

Inhibitory mechanisms of the 14-3-3 dimer

The crystal structure described above shows that the MEK1 and BRAF kinase domains can adopt their mutually inhibited conformations in the absence of any interactions with the 14-3-3 protein. This raises the question of the role of the 14-3-3 dimer in BRAF inhibition. Our structure suggests that, rather than inducing an inactive conformation in the kinase domain, the 14-3-3 maintains the inhibited state by sterically blocking formation of the BRAF kinase domain dimer that is required for BRAF activation[11]. In the cryo-EM structure of the autoinhibited complex, the surface corresponding to the BRAF dimer interface is obstructed by the bound 14-3-3 dimer (Fig. 3a–c). In particular, dimer interface residues H510, D565 and Y566 are all in contact with the 14-3-3 domain. Additionally, the 14-3-3 domain sequesters the CRD domain, which is crucial for Ras-driven activation and membrane recruitment of BRAF. The surface corresponding to the membrane-binding loops of the CRD is largely occluded in the autoinhibited complex (Fig. 3d).

The overall architecture of autoinhibited BRAF is probably shared with both ARAF and CRAF, as key interdomain contacts are highly conserved among these proteins (Extended Data Fig. 6). Consistent with our structural findings, early structure–function studies established a key role for the CRD in maintaining RAF in an autoinhibited state[21,26]. An alanine scanning mutagenesis study of the CRAF CRD domain identified mutations in 11 surface-exposed residues that increased the RAS(G12V)-dependent activation of CRAF, including two that fully activated CRAF in the absence of mutant RAS[27]. All 11 of the corresponding residues in BRAF are located at interdomain contacts in the present structure (Extended Data Fig. 2b). Perhaps most compellingly, the BRAF CRD domain is a hot spot for germline mutations that cause Noonan syndrome and related RASopathies[28]. Altered residues map to sites of contact with 14-3-3 or the BRAF kinase domain in the present structure, providing a structural rationale for their activating effects (Extended Data Fig. 3g).

Structures of active BRAF complexes

The autoinhibited structure described above reveals a clear role for phosphorylation of both S365 and S729 in RAF autoinhibition. To further explore the role of these modifications in RAF regulation, we

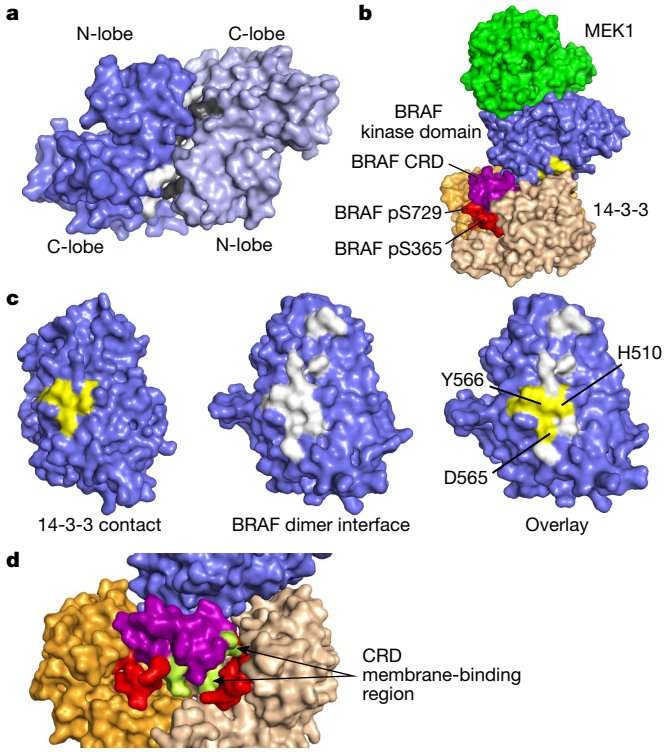


Fig. 3 | The 14-3-3 domain blocks the BRAF dimer interface and occludes the membrane-binding region of the CRD domain. a, Surface representation of an active BRAF kinase domain dimer, with interface residues shaded white for one subunit and grey for the opposite subunit (drawn from PDB entry 4MNE). **b**, Oblique view of the autoinhibited BRAF complex, with BRAF kinase domain residues that contact the α8–α9 loop of the 14-3-3 domain shaded yellow. **c**, Comparison of the 14-3-3 contact (left) and the BRAF dimer interface (centre). The respective surfaces are overlaid on the right, demonstrating that the interaction with 14-3-3 will sterically interfere with the dimerization of BRAF kinase. **d**, Surface view of the complex in the region of the CRD, with the putative membrane-binding loops of the CRD shaded green.

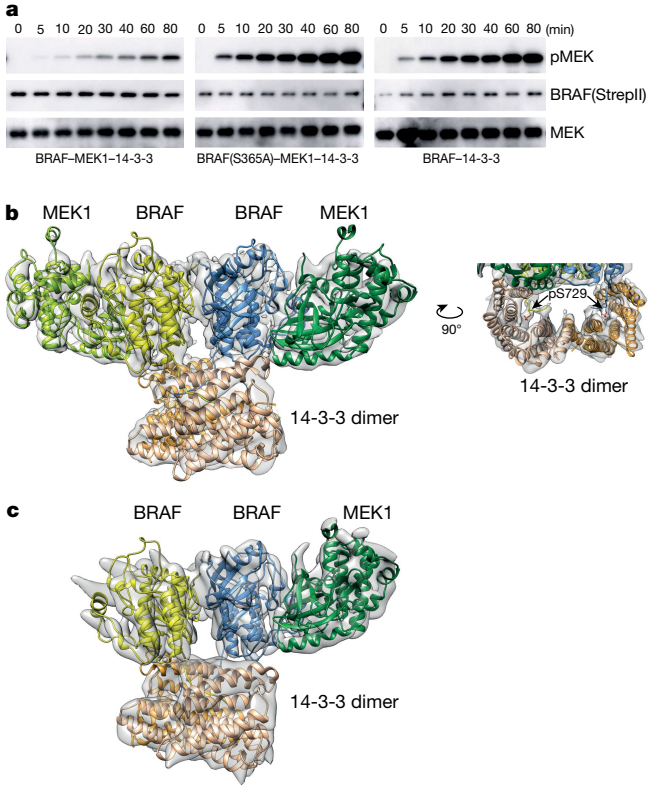


Fig. 4 | Structure and activity of active, dimeric BRAF–MEK1–14-3-3 complexes. a, Kinase activity assays for autoinhibited and active BRAF complexes. Time course of phosphorylation of an exogenous, kinase-dead MEK1 substrate is measured by western blotting for pS218/222 MEK for the autoinhibited wild-type monomer complex (left), the BRAF(S365A) complex (centre), and a wild-type BRAF–14-3-3 complex prepared without co-expression of MEK1 in insect cells (right). Blots for BRAF (anti-strepII) and MEK are provided as loading controls for enzyme and substrate, respectively. For gel source data, see Supplementary Fig. 1. **b**, Cryo-EM structure of the active, dimeric BRAF(S365A)–MEK1–14-3-3 complex. In this active configuration, the 14-3-3 dimer bridges the S729-phosphorylated tails of the back-to-back BRAF dimer (right). Portions of BRAF that are N-terminal to the kinase domain are not visible in the reconstruction. **c**, The same preparations also contained BRAF(S365A)–MEK1–14-3-3 complexes with only one MEK1 subunit. The overall organization of the remaining subunits is very similar, but the 14-3-3 dimer pivots to the side of the missing MEK. See also Extended Data Fig. 8.

prepared the following BRAF variants: BRAF(S365A), with a serine-to-alanine mutation at residue 365; BRAF(S729A); and the double mutant BRAF(S365A/S729A). We then expressed these variants with or without the co-expression of MEK1 in insect cells. Although we obtained soluble, stable BRAF–MEK1–14-3-3 complexes in experiments with BRAF(S365A) (Extended Data Fig. 7a), experiments with the S729A and S365A/S729A variants yielded little BRAF, and it was largely aggregated and did not co-purify with 14-3-3 proteins (data not shown). Size-exclusion chromatography of the BRAF(S365A) sample revealed a broad peak containing BRAF(S365A), MEK1 and the 14-3-3 dimer. Examination of the phosphorylation state of the BRAF(S365A) in this peak revealed near stoichiometric phosphorylation of S729, but little phosphorylation of activation-segment sites T599 and S602 (Extended Data Fig. 1c). Nevertheless, the purified complex was highly active in MEK phosphorylation assays (Fig. 4a, Extended Data Fig. 7a), and cryo-EM imaging of the complex revealed 2D class averages that were consistent with larger, dimeric complexes (Extended Data Fig. 8a).

A three-dimensional (3D) reconstruction, at approximately 5 Å resolution, of the predominant species in this BRAF(S365A)–MEK1–14-3-3 sample revealed an active, back-to-back BRAF kinase dimer, with MEK1 bound to each BRAF kinase domain (Fig. 4b, Extended Data Fig. 8b). A single 14-3-3 dimer bridges the phosphorylated pS729 sites at the C termini of the two BRAF kinase domains. We do not observe interpretable density for BRAF regions preceding the kinase domain, nor for the C terminus beyond S734. We built a model into this cryo-EM map by domain-wise rigid-body fitting of the previously reported active MEK–BRAF kinase domain complex (PDB ID: 4MNE) and the 14-3-3 dimer from the autoinhibited complex described here. The BRAF–MEK kinase domain portion of the structure exhibits the same overall organization as the previous structure, and inspection of the cryo-EM map confirms that the BRAF C-helix is in its inward position, as expected for the active dimer.

Three-dimensional classification of particles from the same set of images enabled the reconstruction of a second particle similar to the one described above, but with only a single MEK1 bound to the BRAF dimer (Fig. 4c). In this 'MEK-lite' complex, the 14-3-3 dimer cants to the side of the missing MEK, assuming a more asymmetric position with respect to the back-to-back BRAF kinase domain dimer.

We also expressed wild-type BRAF alone (without co-expression of MEK1) in both insect and mammalian (HEK293) cells, and obtained soluble BRAF in complex with endogenous 14-3-3 proteins using both expression systems (Extended Data Fig. 7b–f). Quantification of phosphorylation in the elution fractions by mass spectrometry revealed near-stoichiometric phosphorylation of S729 and a high level of S365 phosphorylation in peak fractions, but negligible phosphorylation of both T599 and S602 in BRAF produced by both mammalian and insect cells (Extended Data Fig. 8c–f). The BRAF–14-3-3 complex was highly active in a MEK phosphorylation assay (Fig. 4a, Extended Data Fig. 7b).

Cryo-EM imaging of the mammalian-expressed complex revealed predominant 2D class averages that were consistent with a 14-3-3-bound BRAF dimer, as did imaging of the same sample supplemented with RAF inhibitor GDC-0879 (Extended Data Fig. 8g). We obtained a 3D reconstruction of the inhibitor-bound BRAF–14-3-3 complex at a nominal resolution of 7 Å, which confirmed the dimeric state of the complex (Extended Data Fig. 8h, i). As with the MEK-bound dimer, the BRAF kinase domain forms the expected symmetrical, back-to-back dimer in this structure and we do not observe the N-terminal domains of BRAF. Despite the fact that the 14-3-3 dimer bridges the pS729 sites of the two kinase domains, it adopts a highly asymmetric position with respect to the kinase dimer (Extended Data Fig. 8i). In this skewed position, the 14-3-3 dimer approaches the active-site cleft and intrudes into the MEK-binding region of one BRAF kinase domain, but not the other.

Discussion

In the quiescent state, BRAF, MEK1, and a 14-3-3 dimer form a tightly integrated signalling device. In light of their extensive interactions, we propose that the RAF–MEK–14-3-3 complex—rather than RAF itself—serves as the RAS-activated switch that initiates signalling through the MAP kinase cascade. Phosphorylation of both of the 14-3-3-binding sites and engagement by a 14-3-3 dimer is required for maturation of RAF into its regulated, inactive state. Our structural and biochemical findings suggest that MEK also contributes to the stability of the inactive state of BRAF, but we do not exclude the possibility that RAFs can assemble into an autoinhibited 14-3-3 complex without MEK. The essential role of pS729 in both the autoinhibited and active states of the kinase, and its stoichiometric phosphorylation in our purified complexes, leads us to suggest that this is a structural phosphorylation, rather than a regulatory one. It is noteworthy that, although phosphorylation on T599 and S602 is widely thought to play a crucial role in BRAF activation[29], we find little to no phosphorylation on these sites in active BRAF–14-3-3 dimers. The potential role of activation-loop phosphorylation in RAF regulation merits further study.

The structures described here provide views of RAF in its quiescent and active states and, in light of previous functional dissection of RAF regulation, they outline a model for RAF activation (Extended Data Fig. 9). In the autoinhibited state the RBD is exposed, enabling recruitment of the quiescent complex to the membrane by activated RAS. By contrast, the CRD and its membrane-binding surface is largely buried by interactions with the 14-3-3 dimer and other segments of RAF, suggesting that its 'extraction' upon RAS binding and membrane localization is the key event that promotes the release of the inhibitory position of the 14-3-3 domain. When released from its inhibitory position, the 14-3-3 dimer can rearrange to bridge the pS729 sites in the C-terminal tails of two BRAFs, driving formation of the active BRAF dimer. Once activated, BRAF can phosphorylate MEK, which promotes its release[14]. Steric effects of the 14-3-3 domain could also modulate the affinity for MEK, as evidenced by the asymmetric position assumed by 14-3-3 upon MEK release.

The inactive-state structures described here reveal the bona fide inactive conformation of BRAF, and thereby provide a structural foundation for understanding its activation by mutations in cancer. Oncogenic mutations of V600 and K601 in the inhibitory turn are not compatible with the structural context of these residues, providing a rationale for their activating effect via destabilization of the inhibitory turn (Fig. 2c). Other less common oncogenic BRAF mutations occur in residues that participate directly in the coordination of ATP and its associated divalent cation (Fig. 2d), and they may destabilize the autoinhibited state by weakening interactions with ATP and/or by disrupting interactions of the glycine-rich loop with the inhibitory turn in the activation segment. Many of the same BRAF residues are also altered in RASopathies[28] (Extended Data Fig. 5f). Outside of the kinase domain, CRAF and ARAF contain somatic point mutations in or near the CR2 phosphorylation site in diverse cancers, including lung adenocarcinoma[30]. These mutations eradicate the CR2 14-3-3-binding site, promoting formation of the active RAF dimer—as we observe here with the BRAF(S365A) mutant. The KIAA1549:BRAF truncation/fusion oncoprotein that is found in paediatric low-grade gliomas lacks the entire CR1 and CR2 regions of BRAF, and is therefore constitutively active[31].

The integral nature of the RAF–MEK–14-3-3 switch has important pharmacologic implications. It is well established that certain MEK and RAF inhibitors can stabilize or destabilize their interaction[14,32–34]. However, the notion that the RAF–MEK–14-3-3 complex—which is distinct from the isolated RAF and MEK kinases—may represent a relevant pharmacologic receptor for a broader range of inhibitors has not, to our knowledge, been systematically explored. Perhaps the most perplexing aspect of RAF-inhibitor pharmacology is the paradoxical activation of the MAP kinase pathway by certain RAF kinase inhibitors[4,35,36]. Diverse RAF inhibitors disrupt autoinhibitory interactions of the BRAF kinase with its N-terminal region[37], and some promote dimerization of the isolated BRAF kinase domain[23]. Considering the extensive interactions of BRAF with ATP in the autoinhibited state, we speculate that RAF inhibitors may promote conformational activation by displacing ATP from quiescent RAF. Whether this leads to observed paradoxical pathway activation will in turn depend upon ensuing cellular events—potentially including changes in RAF phosphorylation state, RAS-binding, membrane localization and 14-3-3 rearrangements—and on the potency of a particular agent as an inhibitor of activated RAF dimers.

Many questions regarding RAF regulation remain. The structures described here and the ability to prepare full-length autoregulated and active BRAF will inform and enable detailed mechanistic studies of RAF activation and RAF-inhibitor pharmacology. In the long term, a deeper understanding of RAF regulation should aid in the development of more effective and better-tolerated therapeutics for RAF-driven cancers.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1660-y.

1. Lavoie, H. & Therrien, M. Regulation of RAF protein kinases in ERK signalling. *Nat. Rev. Mol. Cell Biol.* **16**, 281–298 (2015).
2. Simanshu, D. K., Nissley, D. V. & McCormick, F. RAS proteins and their regulators in human disease. *Cell* **170**, 17–33 (2017).
3. Terrell, E. M. & Morrison, D. K. Ras-mediated activation of the Raf family kinases. *Cold Spring Harb. Perspect. Med.* **9**, a033746 (2019).
4. Holderfield, M., Deuker, M. M., McCormick, F. & McMahon, M. Targeting RAF kinases for cancer therapy: BRAF-mutated melanoma and beyond. *Nat. Rev. Cancer* **14**, 455–467 (2014).
5. Sanchez-Vega, F. et al. Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337 (2018).
6. Davies, H. et al. Mutations of the *BRAF* gene in human cancer. *Nature* **417**, 949–954 (2002).
7. Gardino, A. K., Smerdon, S. J. & Yaffe, M. B. Structural determinants of 14-3-3 binding specificities and regulation of subcellular localization of 14-3-3-ligand complexes: a comparison of the X-ray crystal structures of all human 14-3-3 isoforms. *Semin. Cancer Biol.* **16**, 173–182 (2006).
8. Ghosh, S. et al. The cysteine-rich region of raf-1 kinase contains zinc, translocates to liposomes, and is adjacent to a segment that binds GTP-ras. *J. Biol. Chem.* **269**, 10000–10007 (1994).
9. Hekman, M. et al. Associations of B- and C-Raf with cholesterol, phosphatidylserine, and lipid second messengers: preferential binding of Raf to artificial lipid rafts. *J. Biol. Chem.* **277**, 24090–24102 (2002).
10. Leonard, T. A. & Hurley, J. H. Regulation of protein kinases by lipids. *Curr. Opin. Struct. Biol.* **21**, 785–791 (2011).
11. Rajakulendran, T., Sahmi, M., Lefrançois, M., Sicheri, F. & Therrien, M. A dimerization-dependent mechanism drives RAF catalytic activation. *Nature* **461**, 542–545 (2009).
12. Weber, C. K., Slupsky, J. R., Kalmes, H. A. & Rapp, U. R. Active Ras induces heterodimerization of cRaf and BRaf. *Cancer Res.* **61**, 3595–3598 (2001).
13. Rushworth, L. K., Hindley, A. D., O'Neill, E. & Kolch, W. Regulation and role of Raf-1/B-Raf heterodimerization. *Mol. Cell. Biol.* **26**, 2262–2272 (2006).
14. Haling, J. R. et al. Structure of the BRAF–MEK complex reveals a kinase activity independent role for BRAF in MAPK signaling. *Cancer Cell* **26**, 402–413 (2014).

15. Diedrich, B. et al. Discrete cytosolic macromolecular BRAF complexes exhibit distinct activities and composition. *EMBO J.* **36**, 646–663 (2017).
16. Nassar, N. et al. The 2.2 Å crystal structure of the Ras-binding domain of the serine/threonine kinase c-Raf1 in complex with Rap1A and a GTP analogue. *Nature* **375**, 554–560 (1995).
17. Mott, H. R. et al. The solution structure of the Raf-1 cysteine-rich domain: a novel ras and phospholipid binding site. *Proc. Natl Acad. Sci. USA* **93**, 8312–8317 (1996).
18. Wan, P. T. et al. Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell* **116**, 855–867 (2004).
19. Improta-Brears, T., Ghosh, S. & Bell, R. M. Mutational analysis of Raf-1 cysteine rich domain: requirement for a cluster of basic aminoacids for interaction with phosphatidylserine. *Mol. Cell. Biochem.* **198**, 171–178 (1999).
20. Travers, T. et al. Molecular recognition of RAS/RAF complex at the membrane: role of RAF cysteine-rich domain. *Sci. Rep.* **8**, 8461 (2018).
21. Clark, G. J. et al. 14-3-3 ζ negatively regulates Raf-1 activity by interactions with the Raf-1 cysteine-rich domain. *J. Biol. Chem.* **272**, 20990–20993 (1997).
22. Molzan, M. et al. Impaired binding of 14-3-3 to C-RAF in Noonan syndrome suggests new approaches in diseases with increased Ras signaling. *Mol. Cell. Biol.* **30**, 4698–4711 (2010).
23. Thevakumaran, N. et al. Crystal structure of a BRAF kinase domain monomer explains basis for allosteric regulation. *Nat. Struct. Mol. Biol.* **22**, 37–43 (2015).
24. Rauen, K. A. The RASopathies. *Annu. Rev. Genomics Hum. Genet.* **14**, 355–369 (2013).
25. Fischmann, T. O. et al. Crystal structures of MEK1 binary and ternary complexes with nucleotides and inhibitors. *Biochemistry* **48**, 2661–2674 (2009).
26. Michaud, N. R., Fabian, J. R., Mathes, K. D. & Morrison, D. K. 14-3-3 is not essential for Raf-1 function: identification of Raf-1 proteins that are biologically activated in a 14-3-3- and Ras-independent manner. *Mol. Cell. Biol.* **15**, 3390–3397 (1995).
27. Daub, M. et al. The RafC1 cysteine-rich domain contains multiple distinct regulatory epitopes which control Ras-dependent Raf activation. *Mol. Cell. Biol.* **18**, 6698–6710 (1998).
28. Sarkozy, A. et al. Germline *BRAF* mutations in Noonan, LEOPARD, and cardiofaciocutaneous syndromes: molecular diversity and associated phenotypic spectrum. *Hum. Mutat.* **30**, 695–702 (2009).
29. Zhang, B. H. & Guan, K. L. Activation of B-Raf kinase requires phosphorylation of the conserved residues Thr598 and Ser601. *EMBO J.* **19**, 5429–5439 (2000).
30. Imielinski, M. et al. Oncogenic and sorafenib-sensitive *ARAF* mutations in lung adenocarcinoma. *J. Clin. Invest.* **124**, 1582–1586 (2014).
31. Pfister, S. et al. *BRAF* gene duplication constitutes a mechanism of MAPK pathway activation in low-grade astrocytomas. *J. Clin. Invest.* **118**, 1739–1749 (2008).
32. Hatzivassiliou, G. et al. Mechanism of MEK inhibition determines efficacy in mutant KRAS- versus BRAF-driven cancers. *Nature* **501**, 232–236 (2013).
33. Ishii, N. et al. Enhanced inhibition of ERK signaling by a novel allosteric MEK inhibitor, CH5126766, that suppresses feedback reactivation of RAF activity. *Cancer Res.* **73**, 4050–4060 (2013).
34. Lito, P. et al. Disruption of CRAF-mediated MEK activation is required for effective MEK inhibition in KRAS mutant tumors. *Cancer Cell* **25**, 697–710 (2014).
35. Karoulia, Z., Gavathiotis, E. & Poulikakos, P. I. New perspectives for targeting RAF kinase in human cancer. *Nat. Rev. Cancer* **17**, 676–691 (2017).
36. Hymowitz, S. G. & Malek, S. Targeting the MAPK pathway in RAS mutant cancers. *Cold Spring Harb. Perspect. Med.* **8**, a031492 (2018).
37. Jin, T. et al. RAF inhibitors promote RAS–RAF interaction by allosterically disrupting RAF autoinhibition. *Nat. Commun.* **8**, 1211 (2017).

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Cell lines

The insect (Sf9 and Hi5) and mammalian (Expi293F) cell lines used for protein production were obtained from Thermo Fisher Scientific, and tested negative for mycoplasma contamination.

Preparation of BRAF–14-3-3 complexes from Sf9 insect cells

Recombinant baculovirus expressing full-length human BRAF with an N-terminal His_8-tag and a C-terminal StrepII tag was prepared using baculoviral transfer vector pAc8. Recombinant baculovirus expressing the variant BRAF(S365A) was produced in the same manner. For protein production using the baculovirus/insect cell expression system, litre-scale cultures of Sf9 cells (4 l total for a typical preparation) were infected with high-titre viral stocks expressing wild-type or mutant BRAF (1% of final culture volume). Cells were collected 65–72 h post-infection, lysed in lysis buffer (50 mM Tris pH 7.4, 150 mM NaCl, 2 mM $MgCl_2$, 0.5 mM TCEP, 50 μM ATP-γ-S and protease inhibitor cocktail (Thermo Fisher Scientific), and applied to Ni-NTA agarose beads (Qiagen). After washing with Buffer A supplemented with 20 mM imidazole (Buffer A contains 50 mM Tris pH 7.4, 150 mM NaCl, 2 mM $MgCl_2$, 0.5 mM TCEP, 10 μM ATP-γ-S), bound proteins were eluted with Buffer A supplemented with 500 mM imidazole and adjusted to pH 8.0. Eluent was applied to a prepacked StrepTrap HP column (GE Healthcare Life Sciences) and washed with Buffer A adjusted to pH 8.0 (Buffer A′). Bound proteins were eluted with Buffer A′ supplemented with 10 mM desthiobiotin. The eluted complex was concentrated to approximately 2 mg ml^{-1} using an Amicon Ultra concentrator (50 MWCO, Millipore) and further purified by size-exclusion chromatography (SEC) on a Superdex 200 Increase 10/300 column or Superose 6 increase 10/300 (GE Healthcare Life Sciences) in Buffer A. Analysis of the purified samples by SDS–PAGE revealed the baculovirus-expressed BRAF co-purified at a high stoichiometry with insect-cell-derived 14-3-3ε and 14-3-3ζ.

Preparation of BRAF–MEK1–14-3-3 complexes from Sf9 insect cells

Recombinant baculovirus expressing full-length human MEK1 (either wild-type MEK1 or MEK1(S218A/S222A)) fused with an N-terminal His_6-tag was prepared using baculoviral transfer vector pAc8. BRAF–MEK1–14-3-3 complexes (either wild-type or with the desired BRAF and/or MEK mutants) were prepared by co-expression in insect cells using separate baculoviruses for MEK1 and BRAF. Litre-scale cultures (4 l total for a typical preparation) were co-infected with high-titre viral stocks expressing the desired BRAF and MEK1 variants at a 1:1.5 ratio (by volume, 1% culture volumes of BRAF virus, 1.5% of MEK1), and cells were collected by centrifugation 65–72 h post-infection. BRAF–MEK1–14-3-3 complexes were purified from cell pellets as described above for BRAF–14-3-3 complexes, but all buffers were supplemented with MEK inhibitor GDC-0623 to a final concentration of 2 μM. As with BRAF alone, co-expressed BRAF and MEK1 co-purified with insect-cell-derived 14-3-3ε and 14-3-3ζ.

Preparation of the BRAF–14-3-3 complex from HEK293 mammalian cells

Full-length human BRAF bearing an N-terminal His_8-tag and a C-terminal StrepII tag was cloned into pcDNA 5/FRT/TO vector. For protein production, litre-scale suspension cultures (2 l total for a typical preparation) of HEK293 cells (Expi293F) were transfected using the Expi293 expression system according to the manufacturer's protocol (Thermo Fisher Scientific). Cells were collected by centrifugation 48–60 h post-transfection. BRAF–14-3-3 complexes were purified from mammalian cell pellets as described above for isolation BRAF–14-3-3 from insect cells. Analysis of the purified sample by SDS–PAGE revealed that BRAF co-purified with mammalian-cell-derived 14-3-3 isoforms.

Preparation of Spycatcher–MEK1 from Hi5 insect cells

We prepared kinase-dead MEK1 (fused to Spycatcher to alter its electrophoretic mobility) for use as a substrate in in vitro BRAF-activity assays. Recombinant baculovirus encoding full-length human MEK1(D190N) bearing an N-terminal His_6-tag for purification and a C-terminal Spy-tag was prepared using baculoviral transfer vector pAc8. For protein production, litre-scale cultures (2 l total for a typical preparation) were infected with high-titre viral stocks expressing Spy-tagged MEK1(D190N) (1% of final culture volume). Cells were collected 55–65 h post-infection, lysed in MEK lysis buffer (50 mM HEPES pH 7.4, 150 mM NaCl, 2 mM $MgCl_2$, 0.5 mM TCEP) and applied to Ni-NTA agarose beads (Qiagen). After washing with MEK lysis buffer supplemented with 20 mM imidazole, bound proteins were eluted with lysis buffer supplemented with 500 mM imidazole and adjusted to pH 7.4. To ensure that MEK1 was not phosphorylated, eluted protein was treated with lambda phosphatase overnight at 4 °C before further purification by SEC on a Superdex 75 Increase 10/300 column in SEC Buffer (50 mM Tris pH 8.0, 150 mM NaCl, 2 mM $MgCl_2$, 1 mM TCEP). Pooled SEC fractions containing Spy-tagged MEK1(D190N) were incubated with Spycatcher protein for covalent linkage, as described previously[38]. Analysis of the purified Spycatcher-MEK1(D190N) protein by SDS–PAGE confirmed that it migrated as expected for a protein of approximately 55 kDa. Mass spectrometry and western blotting with pMEK1/2 antibody confirmed little or no phosphorylation on the MEK1 activation loop (S218 and S222).

Size-exclusion chromatography with multiangle light scattering

The BRAF–MEK1–14-3-3 complex (prepared with MEK1(S218A/S222A)) was applied to a Superdex 200 10/300 GL column (GE Healthcare) in 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 2 mM $MgCl_2$, 0.5 mM TCEP, 10 μM ATP-γ-S, 2 μM GDC-0623. In-line multi-angle light scattering analysis was performed with an OptiLab rEX refractive index detector followed by a miniDAWN TREOS light scattering detector, and data were analysed with ASTRA (Wyatt Technology).

Kinase-activity assay

BRAF activity in SEC elution fractions was measured by diluting an aliquot of each fraction fivefold, and adding 1 μl of the diluted sample to 14 μl of a reaction mixture containing assay buffer (50 mM Tris pH 7.5, 150 mM NaCl, 10 mM $MgCl_2$, 0.5 mM TCEP, and 1 mM sodium vanadate) supplemented with 2.67 μM Spycatcher–MEK1(D190N) as a substrate. Kinase reactions were started by addition of 5 μl of 4 mM ATP in assay buffer. After incubation for 20 min at 25 °C, reactions were stopped by addition of SDS–PAGE sample buffer and heating to 95 °C. Reaction products were resolved on a Novex 12% Tris-Glycine Midi gel (Invitrogen), and subsequently western-blotted with anti-phosphoMEK1/2(S218/222) antibody (Cell Signaling Technology) to detect phosphorylation of the Spycatcher–MEK substrate (approximately 55 kDa).

Time-course kinase assays of BRAF–MEK1–14-3-3, BRAF(S365A)–MEK1–14-3-3 and BRAF–14-3-3 samples were performed using the same reaction buffer and substrate concentrations described above, but each sample was diluted to 100 nM, and 20 μl was used in a final reaction volume of 200 μl (final enzyme concentration was 10 nM in the reaction mixture). After initiating the assay by addition of ATP, 20 μl aliquots were removed from the reaction at the designated time points and stopped by mixing with an equal volume of 5X SDS–PAGE sample buffer and heating to 95 °C. Reaction products were analysed by SDS–PAGE and western blotting for phosphoMEK1/2 as described above. The BRAF–MEK1–14-3-3 complexes were prepared with MEK1(S218A/S222A).

Pull-down assay for 14-3-3 association

Recombinant baculoviruses expressing full-length human 14-3-3β/α, 14-3-3γ, 14-3-3δ/ζ or 14-3-3ε and bearing an N-terminal Flag-tag were prepared using baculoviral transfer vector pAc8. BRAF and MEK1 were co-expressed with each of the four different human 14-3-3 isoforms in insect cells by co-infection. Co-infected Sf9 cells (100 ml) were collected 55–65 h post-infection, lysed in lysis buffer, and then parallel aliquots of clarified lysate were applied to Strep-TactinXT magnetic beads (IBA GmbH) or Anti-DYKDDDDK magnetic Agarose (anti-Flag, Pierce). After washing beads with lysis buffer, bound proteins were eluted with SDS–PAGE sample buffer and resolved on 8% Bis-Tris PAGE gels. Parallel gels were western-blotted using anti-14-3-3 (pan) and anti-Flag antibodies (Cell Signaling Technology).

Cryo-EM data acquisition and processing

The MEK1(S218A/S222A) variant was used to prepare all BRAF–MEK1–14-3-3 complexes for structural analysis. BRAF–MEK1–14-3-3 complex in SEC buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 2 mM $MgCl_2$, 0.5 mM TCEP, 10 μM ATP-γ-S, 2 μM GDC-0623) was applied to glow-discharged holey carbon grids (Quantifoil R1.2/1.3, 400 mesh) and vitrified using a FEI Vitrobot Mark IV. Frozen hydrated samples were imaged on an FEI Titan Krios at 300 kV with a Gatan Quantum Image Filter with K2 Summit direct detection camera in super-resolution mode with a total exposure dose of around 50 electrons. Thirty-five frames per movie were collected at a magnification of 130,000×, corresponding to 0.53 Å per pixel. In total, 8,440 micrographs were collected at defocus values ranging from −1.8 to −3.3 μm from two data collections of 4,097 and 4,343 images (of which 1,628 images in the later collection were tilted by 25° in an attempt to increase observed orientations). The movie frames were motion-corrected and dose-weighted by MotionCor2[39], downsampled to 1.06 Å per pixel and contrast transfer function (CTF) parameters were estimated by CTFFIND4[40]. Particle picking was carried out using crYOLO[41] and template-based particle picking within Relion[42] giving 3,531,955 initial particles. Following successive rounds of 2D and 3D classification 427,592 particles were selected. An additional two rounds of 3D classification led to the final reconstruction of 4.1 Å from 165,298 particles. Per particle motion correction 'particle polishing' alongside per particle CTF refinement was trialled with no improvement in resolution or map quality. Maps used for figures were filtered according to local resolution with b-factor sharpening within Relion[42]. Models were fit into the map using Coot[43] and further refined with PHENIX[44] and REFMAC5[45]. Statistics for the final refinement are presented in Extended Data Table 1.

The BRAF(S365A)–MEK1–14-3-3 complex in SEC buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 2 mM $MgCl_2$, 0.5 mM TCEP, 10 μM ATP-γ-S, 2 μM GDC-0623) was applied to glow-discharged holey carbon grids (Quantifoil R1.2/1.3, 400 mesh) and vitrified using a Leica EM GP. Frozen hydrated samples were imaged on an FEI Talos Arctica at 200 kV with K3 Summit direct detection camera in counting mode with a total exposure dose of around 50 electrons. Forty-two frames per movie were collected at a magnification of 36,000×, corresponding to 1.11 Å per pixel. Micrographs (3,146) were collected at defocus values ranging from −2.0 to −3.0 μm. Initial particle picking was carried with crYOLO[41] and ab-initio models were generated in cryoSPARC[46] showing two distinct classes: one showing back-to-back BRAF kinase dimer, with MEK1 bound to each BRAF kinase domain; the other having only a single MEK1 bound (MEK-lite). All following steps were carried out within Relion[42]. Reference-based picking resulted in 2,008,323 particles. Following 2D classification 1,441,851 particles were subjected to a guided 3D classification using single copies of the dimeric and MEK-lite as reference models. Dimer particles (705,222) were then subjected to a further round of standard 3D classification leaving 425,135 particles. After Bayesian polishing and 3D refinement this resulted in a reconstruction of 4.9 Å. In addition, 736,629 particles were identified as MEK-lite and, after further 3D classification, 595,672 particles were subjected to Bayesian polishing and 3D refinement resulting in a 5.7 Å reconstruction. Models were built for both reconstructions by rigid-body fitting the BRAF and MEK1 kinase domains from PDB entry 4MNE using Coot[43]. The 14-3-3 domain was modelled by rigid-body fitting of the insect-cell 14-3-3 ζ-domain from the autoinhibited BRAF–MEK1–14-3-3 structure described here; each subunit of the 14-3-3 dimer was fit independently. C-terminal pS729 tails were manually built for each structure, with reliance on the autoinhibited structure for placement of pS729. For both reconstructions, cryo-EM maps were deposited in the Electron Microscopy Data Bank and polyalanine models were deposited in the Protein Data Bank. Data collection and image processing statistics for both structures are presented in Extended Data Table 1.

The mammalian-cell-produced BRAF–14-3-3 complex in SEC buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 2 mM $MgCl_2$, 0.5 mM TCEP, 10 μM ATP-γ-S, 2 μM GDC-0623) with or without 1 μM GDC-0879 was applied to glow-discharged holey carbon grids (Quantifoil R1.2/1.3, 400 mesh) and vitrified using a Leica EM GP. Frozen hydrated samples were imaged on an FEI Titan Krios at 300 kV with a Gatan Quantum Image Filter with K3 Summit direct detection camera in counting mode with a total exposure dose of around 70 electrons. Fifty frames per movie were collected at a magnification of 105,000×, corresponding to 0.85 Å per pixel. Micrographs (4,002 and 4,418 per sample, with and without 1 μM GDC-0879) were collected at defocus values ranging from −1.7 to −2.7 μm. The movies were downsampled to 1.7 Å and particle picking was carried out on the GDC-0879 sample with crYOLO[41] giving 365,083 particles. After two rounds of 2D classification, 234,539 particles remained. Two further rounds of 3D classification within Relion[42], using an initial model derived from cryoSPARC[46], resulted in 66,215 particles leading to a 6.8 Å reconstruction after 3D refinement. The BRAF–14-3-3 model was constructed by rigid-body fitting the BRAF and MEK1 kinase domains from PDB entry 4MNE using Coot[43]. The 14-3-3 domain was modelled by rigid-body fitting of the human 14-3-3 ζ-domain (PDB ID: 3NKX). Each subunit of the 14-3-3 dimer was fit independently, and the C-terminal pS729 tails were manually built. The cryo-EM map has been deposited in the Electron Microscopy Data Bank and the polyalanine model has been deposited in the Protein Data Bank. Data collection and image processing statistics for this structure are presented in Extended Data Table 1.

Expression and purification of the BRAF–MEK kinase domain complex

For insect-cell expression of the BRAF kinase domain in complex with MEK1, two recombinant baculovirus species were used. The first was prepared using baculoviral transfer vector pFastBac Dual and encoded the BRAF kinase domain (BRAF residues 445–723, fused to an N-terminal His_6-tag and a C-terminal chitin-binding domain) and human chaperone CDC37. The second baculovirus encoded full-length human MEK1(S218A/S222A), as described above. For protein production, litre-scale suspension cultures of Sf9 cells were co-infected with both viruses. Cells were collected 60–66 h post-infection and resuspended in lysis buffer (50 mM Tris pH 8.0, 250 mM NaCl, 5% glycerol and 20 mM imidazole, 1 mM TCEP) with protease inhibitor cocktail (Roche). Resuspended cells were disrupted by sonication on wet ice, and the lysate was clarified by ultracentrifugation at 40,000 r.p.m. for two hours. Clarified lysate was batch-bound to Ni-NTA beads and washed extensively with binding buffer before elution with elution buffer (50 mM Tris pH 8.0, 250 mM NaCl, 250 mM imidazole, 1 mM TCEP). The elution fractions were pooled and treated with 1:1,000 molar ratio of TEV protease and 100 mM β-mercaptoethanesulfonic acid (MESNA) overnight to cleave N-terminal and C-terminal tags, and further purified by SEC (Superdex 200 10/300, GE Healthcare) in storage buffer (50 mM HEPES pH 7.5, 150 mM NaCl, 1 mM TCEP). The fractions were analysed by SDS–PAGE, and fractions corresponding to the BRAF–MEK1 kinase domain complex were pooled, concentrated to 8 mg ml^{-1}, and flash-frozen.

BRAF–MEK kinase domain crystallization and structure determination

For crystallization, an aliquot of the BRAF–MEK1(S218A/S222A) kinase complex was incubated with 5 mM $MgCl_2$, 2 mM adenosine 5′-(β,γ-imido)triphosphate (AMP-PNP), and 0.2 mM GDC-0623 in storage buffer at 4 °C overnight. Rod-shaped crystals suitable for structure determination were obtained by vapour diffusion in hanging drops using a reservoir solution consisting of 100 mM Bis-Tris pH 6.5, 200 mM ammonium sulfate, and 22% PEG 3350 at room temperature. Crystals were collected and flash-frozen in liquid nitrogen using additional 20% glycerol as a cryoprotectant. X-ray diffraction data were collected at 100 K using NE-CAT beamline ID-24-C at the Advance Photon Source, Argonne National Laboratory, at a wavelength of 0.979 Å. Data were integrated and merged using XDS[47] and scaled using Aimless in the CCP4 suite[48]. The structure was phased by molecular replacement in PHASER[49] using the relevant domains of the autoinhibited cryo-EM structure and PDB entry 4MNE as initial search models. GDC-0623 was placed into positive density in an initial $F_o - F_c$ map and included in subsequent rounds of refinement using PHENIX.REFINE[50]. Successive manual refinement was performed using Coot[43]. The structure was refined to R_{work}/R_{free} values of 0.22/0.25 at a resolution of 2.58 Å. Data collection and refinement statistics are presented in Extended Data Fig. 5a.

Mass spectrometry analysis

BRAF complexes were digested separately with trypsin and Lys-C, desalted by C18, dried by vacuum centrifugation, and analysed in triplicate by capillary electrophoresis coupled to mass spectrometry (CE–MS) using a ZipChip autosampler and CE instrument (908 Devices) interfaced to a QExactive HF mass spectrometer (Thermo Fisher Scientific). Peptides were loaded for 20 s on an HR chip and electrophoresis was performed at 700 V cm^{-1} for 10 min, with pressure assist activated at 1 min. To identify BRAF phosphopeptides, digests were analysed by data-dependent tandem mass spectrometry (MS/MS). The five most abundant ions in each MS scan (60K resolution) were subjected to MS/MS (15K resolution, 30% collision energy). Dynamic exclusion was activated with a repeat count of 1 and an exclusion duration of 6 s. MS/MS spectra were converted to .mgf format using multiplierz software[51,52], and searched against a database of BRAF using Mascot 2.6.1. Search parameters specified trypsin or Lys-C specificity with up to two missed cleavages; variable oxidation of methionine; variable phosphorylation of serine, threonine or tyrosine; fixed carbamidomethylation of cysteine; and precursor and product ion tolerances of 10 ppm and 25 mmu, respectively. Identified sites of phosphorylation were confirmed using mzStudio software[53]. In experiments to determine the stoichiometry of phosphorylation, digests were analysed by CE–MS (MS1 scans with 15K resolution), with precursor peak areas used for quantification according to the following equations:

$$\%\ \text{Phosphorylation} = \frac{A_{\text{Corr.P-pep}}}{A_{\text{Tot}}}$$

$$A_{\text{Tot}} = A_{\text{Corr.P-pep}} + A_{\text{NonP-pep}}$$

$$A_{\text{Corr.P-pep}} = A_{\text{P-pep}} \times \text{Corr. factor}$$

where $A_{\text{Corr.P-pep}}$ is the area of the phosphopeptide corrected for differences in ionization efficiency due to phosphorylation, A_{Tot} is the total peptide peak area, $A_{\text{NonP-pep}}$ is the peak area of the unphosphorylated peptide, $A_{\text{P-pep}}$ is the uncorrected peak area of the phosphopeptide, and Corr. factor is the correction factor for ionization efficiency. Correction factors for ionization efficiency were determined in separate experiments using either BRAF protein digests or synthetic BRAF peptide standards. Peptides, with or without treatment with alkaline phosphatase (pptase), were analysed in triplicate by CE–MS as described above. After normalizing for loading amounts (using non-phosphorylatable BRAF peptides VFLPNK and LIDIAR for BRAF digests or a spiked standard peptide for BRAF synthetics), correction factors were calculated according to:

$$\text{Corr. factor} = \frac{A_{\text{NonP-pep(+pptase)}} - A_{\text{NonP-pep(-pptase)}}}{A_{\text{P-pep(-pptase)}}}$$

where $A_{\text{NonP-pep(+pptase)}}$ is the area of the non-phosphorylated peptide after phosphatase treatment, $A_{\text{NonP-pep(-pptase)}}$ is the area of the non-phosphorylated peptide without phosphatase treatment, and $A_{\text{P-pep(-pptase)}}$ is the area of the phosphorylated peptide without phosphatase treatment.

Because discovery experiments did not detect phosphorylation of MEK1 activation loop sites 218/222 (peptide $_{206}$LCDFGVSGQLIDSMANSFVGTR$_{227}$) we estimated an upper bound of phosphorylation of these residues by analysing digests with or without pptase treatment as above. In these experiments, we used targeted selected ion monitoring scans of $_{206}$LCDFGVSGQLIDSMANSFVGTR$_{227}$ and normalization peptides, and then calculated MEK1 activation loop phosphorylation as:

$$\%\text{Phosphorylation} = 1 - \frac{A_{\text{NonP-MEK(-pptase)}}}{A_{\text{NonP-MEK(+pptase)}}}$$

where $A_{\text{NonP-MEK(-pptase)}}$ corresponds to the peak area of the unphosphorylated MEK peptide without phosphatase treatment and $A_{\text{NonP-MEK(+pptase)}}$ corresponds to the peak area of the unphosphorylated MEK peptide after phosphatase treatment. Data analysis and peak integration were performed using mzStudio software[53].To identify 14-3-3 proteins, MS/MS spectra from data-dependent CE–MS analyses of trypsin and Lys-C digested BRAF complexes were converted to .mgf format using multiplierz software[51,52], and searched against a forward-reverse human protein database (uniprot) with Mascot 2.6.1 (using the same search parameters as described above). Data were filtered to a false discovery rate of around 1%, and peptide sequences mapped to genes using the multiplierz pep2gene tool[52].

Sequence alignments and software

For Extended Data Figs. 2a, 6, sequences were aligned using ClustalW and figures were prepared with ESPript 3.0[54]. Structural biology applications used in this project were compiled and configured by SBGrid[55].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Three-dimensional cryo-EM density maps have been deposited in the Electron Microscopy Data Bank (EMDB) with accession codes EMD-0541, EMD-20550, EMD-20552 and EMD-20551. Atomic coordinates corresponding to these cryo-EM reconstructions have been deposited in the Protein Data Bank (PDB) with accession codes 6NYB, 6Q0J, 6Q0T and 6Q0K. Structure factors and atomic coordinates for the BRAF kinase domain–MEK1 crystal structure have been deposited in the PDB with accession code 6PP9.

38. Zakeri, B. et al. Peptide tag forming a rapid covalent bond to a protein, through engineering a bacterial adhesin. *Proc. Natl Acad. Sci. USA* **109**, E690–E697 (2012).
39. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
40. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
41. Wagner, T. et al. SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. *Commun Biol.* **2**, 218 (2019).
42. Zivanov, J., Nakane, T., Forsberg, B. O., Kimanius, D., Hagen, W. J., Lindahl, E. & Scheres, S. H. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7**, e42166 (2018).

43. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
44. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
45. Murshudov, G. N. et al. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D* **67**, 355–367 (2011).
46. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
47. Kabsch, W. XDS. *Acta Crystallogr. D* **66**, 125–132 (2010).
48. Potterton, E., Briggs, P., Turkenburg, M. & Dodson, E. A graphical user interface to the CCP4 program suite. *Acta Crystallogr. D* **59**, 1131–1137 (2003).
49. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr*. **40**, 658–674 (2007).
50. Afonine, P. V. et al. Towards automated crystallographic structure refinement with phenix. refine. *Acta Crystallogr. D* **68**, 352–367 (2012).
51. Alexander, W. M., Ficarro, S. B., Adelmant, G. & Marto, J. A. multiplierz v2.0: A Python-based ecosystem for shared access and analysis of native mass spectrometry data. *Proteomics* **17**, 1700091 (2017).
52. Parikh, J. R. et al. multiplierz: an extensible API based desktop environment for proteomics data analysis. *BMC Bioinformatics* **10**, 364 (2009).
53. Ficarro, S. B., Alexander, W. M. & Marto, J. A. Mzstudio: A dynamic digital canvas for user-driven interrogation of mass spectrometry data. *Proteomes* **5**, 20 (2017).
54. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320–W324 (2014).
55. Morin, A., Eisenbraun, B., Key, J., Sanschagrin, P. C., Timony, M. A., Ottaviano, M. & Sliz, P. Collaboration gets the most out of software. *eLife* **2**, e01456 (2013).

Acknowledgements This work was supported in part by the PLGA fund at the Pediatric Brain Tumor Foundation, by Novartis Institutes for Biomedical Research, and by National Institutes of Health (NIH) grants P50 CA165962 (M.J.E.), PO1 CA154303 (M.J.E.) and R50 CA221830 (E.P.). Cryo-EM imaging for the autoinhibited structure was carried out at the University of Massachusetts Medical School Cryo-EM Core Facility, and we thank C. Xu and K. Song for their guidance in image acquisition. Active-state structures were imaged at the Harvard Cryo-EM Center for Structural Biology. Diffraction data for the BRAF/MEK1 kinase complex crystal structure were recorded at beamline 24-ID-C of the Northeast Collaborative Access Team at the Advanced Photon Source, Argonne National Laboratory, which is supported in part by the Department of Energy and NIH grant GM124165. We thank M. Kostic for critical reading and editing of the manuscript.

Author contributions E.P. expressed, purified and biochemically characterized all BRAF–MEK1–14-3-3 and BRAF–14-3-3 complexes. E.P. and H.J. prepared the samples and collected electron microscopy data. S.R., together with H.J. and M.J.E., processed the electron microscopy data to obtain the 3D reconstructions. E.P., S.R., B.-W.K., K.L., H.J. and M.J.E. built and analysed cryo-EM models. K.L. and G.G.-D.P. determined the BRAF–MEK1 kinase domain crystal structure. S.B.F. and J.A.M. were responsible for the mass spectrometry experiments. H.S. carried out preliminary expression and purification studies for the complex. H.J. and M.J.E. directed the project, and M.J.E. drafted the manuscript with input from all authors.

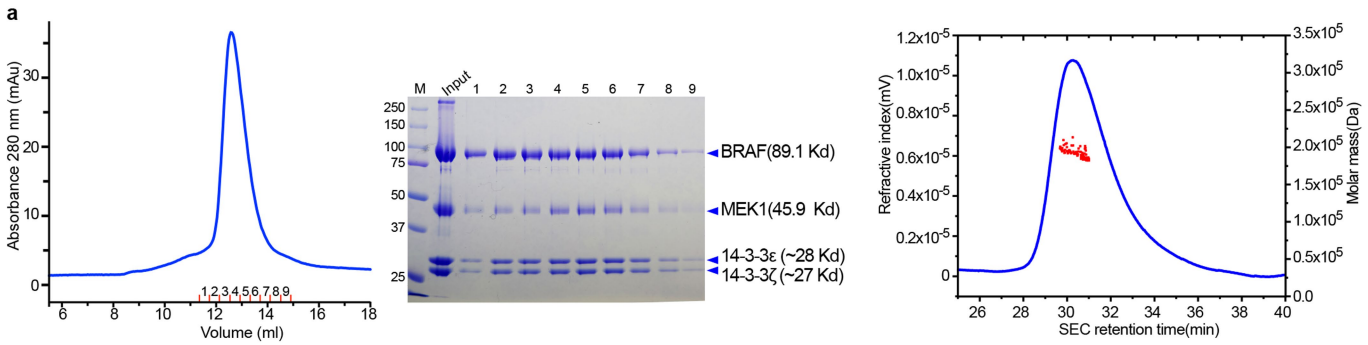
Competing interests This work was supported in part by a sponsored research agreement from Novartis Institutes for Biomedical Research (NIBR) and M.J.E. has been a consultant to NIBR. J.A.M. serves on the Scientific Advisory Board of 908 Devices.

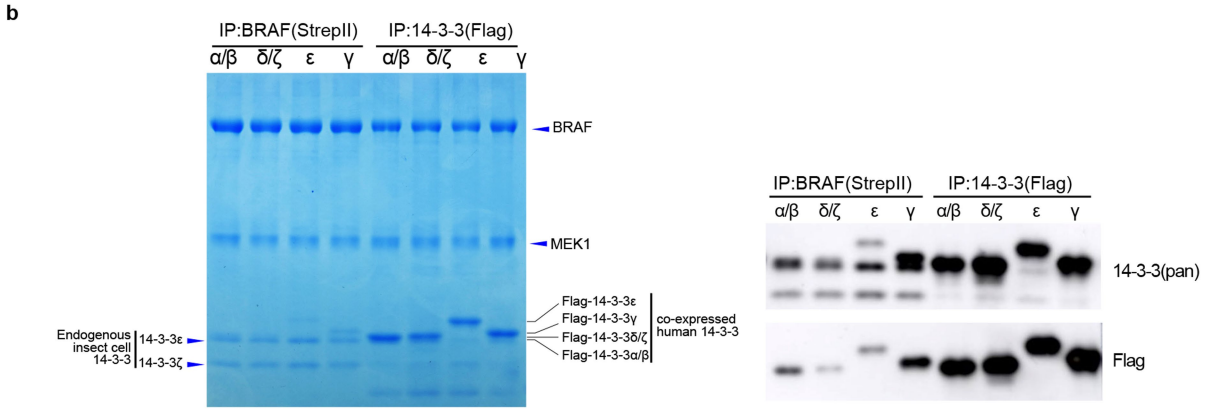
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1660-y.

Correspondence and requests for materials should be addressed to H.J. or M.J.E.

Peer review information *Nature* thanks Frank Sicheri and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

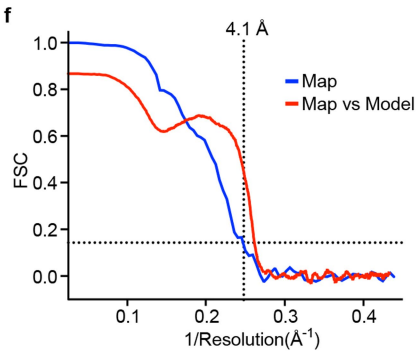


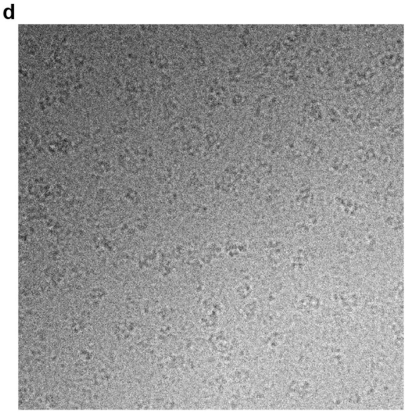


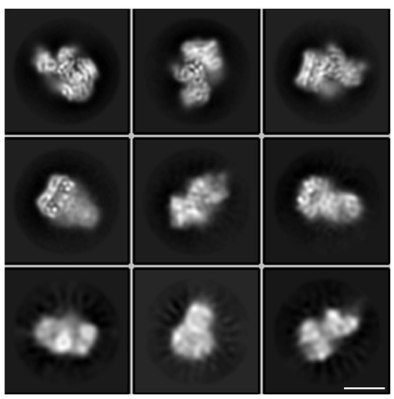
c

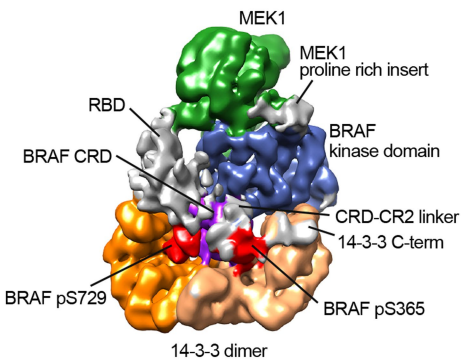
Protein	Site	BRAF–MEK1–14-3-3 % Phosphorylation	BRAF(365A)–MEK1–14-3-3 % Phosphorylation
BRAF	Ser 365	95%	N/A
	Ser 446	83%	71%
	Thr 599	1%	0%
	Ser 602	1%	9%
	Ser 605	11%	48%
	Ser 729	99%	99%
MEK1	Ser 218/222	≤11%*	N/A

*Phosphosite not detected; upper bound based on increase in signal for non-phosphorylated species after phosphatase treatment.



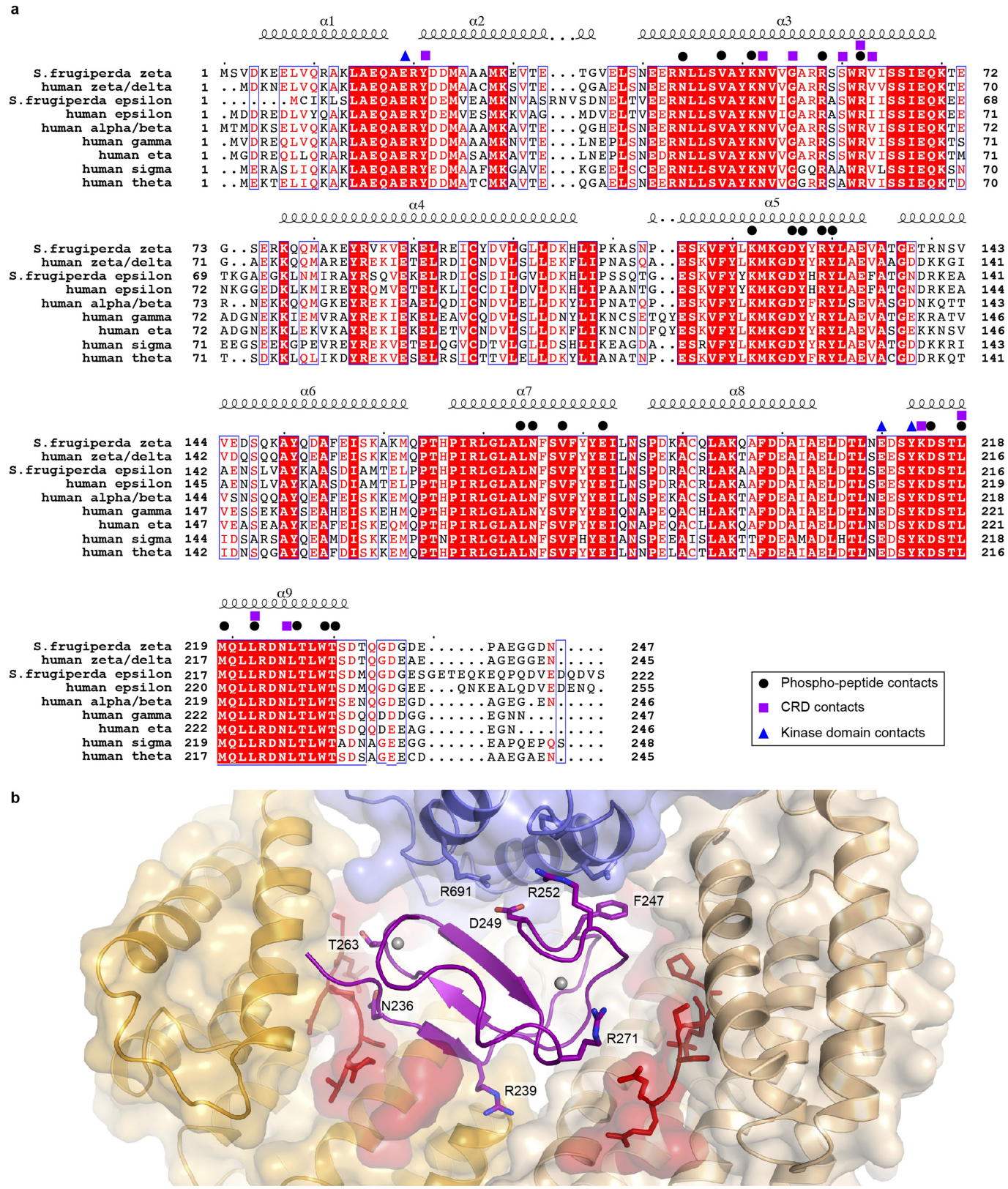






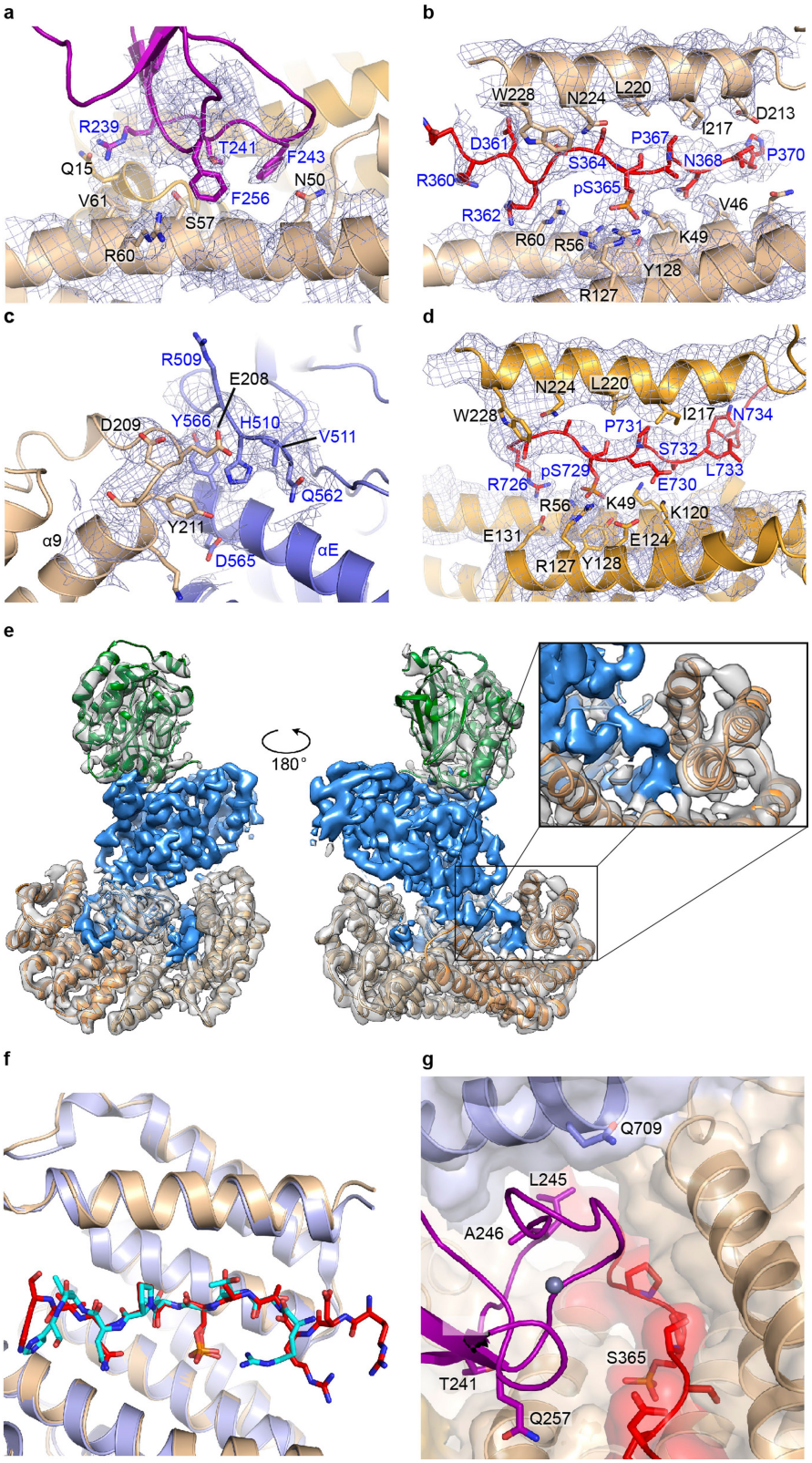
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Biochemical characterization of purified BRAF complexes and cryo-EM analysis of the autoinhibited BRAF–MEK1–14-3-3 complex. a, The full-length, autoinhibited BRAF–MEK1–14-3-3ε,ζ complex used for cryo-EM structure determination. Left, the elution profile from SEC on a Superdex 200 column; centre, Coomassie-stained SDS–PAGE analysis of elution fractions; right, analysis by size-exclusion chromatography with multi-angle light scattering. A molar mass of 196 kDa was indicated; the calculated molecular weight of the complex is 192 kDa. **b**, Analysis of co-expression of human 14-3-3 isoforms with BRAF and MEK1 in insect cells. Left, Strep-tagged BRAF, MEK1(S218A/S222A) and the indicated Flag-tagged human 14-3-3 isoforms were co-expressed in Sf9 cells and BRAF–MEK1–14-3-3 complexes were affinity-isolated from clarified lysates with either Strep-TactinXT (left four lanes) or anti-Flag (right four lanes) magnetic beads. Right, parallel gels were blotted with an anti-14-3-3 antibody that recognizes all 14-3-3 isoforms (top blot) or with an anti-Flag antibody (bottom blot). Note that even in the presence of robust overexpression of these human isoforms, BRAF preferentially associated with the endogenous insect-cell 14-3-3 proteins (as seen in the Strep-TactinXT-precipitated lanes of the Coomassie-stained gel). **c**, Mass-spectrometry-based quantification of selected phosphorylation sites in complexes with wild-type BRAF and with BRAF(S365A) purified for structural analysis. Note that the BRAF(S365A) complex was prepared with MEK1(S218A/S222A), whereas the wild-type BRAF complex used in this analysis contained wild-type MEK1. **d**, Portion of a representative micrograph used for reconstruction of the autoinhibited BRAF–MEK1–14-3-3 complex. **e**, Representative 2D class averages for reconstruction of the autoinhibited BRAF–MEK1–14-3-3 complex. Scale bar, 10 nm. **f**, Fourier shell correlation (FSC) curves for the reconstruction. The horizontal line indicates a correlation of 0.143; the FSC curve for two half-maps (blue) crosses this threshold at a resolution of 4.1Å. A correlation curve for the map versus the atomic model is plotted in red. **g**, The cryo-EM map of the autoinhibited BRAF–MEK1–14-3-3 complex filtered to 5 Å resolution and contoured at a lower level to reveal weaker density corresponding to the RBD domain. The map surface is coloured by domain as in Fig. 1. Unassigned densities (grey) can be ascribed to the RBD domain and other poorly structured elements as indicated. For gel source data, see Supplementary Fig. 1. Experiments in **a** and **b** were repeated at least twice with similar results. Imaging experiments in **d** and **e** were repeated four times with similar results.



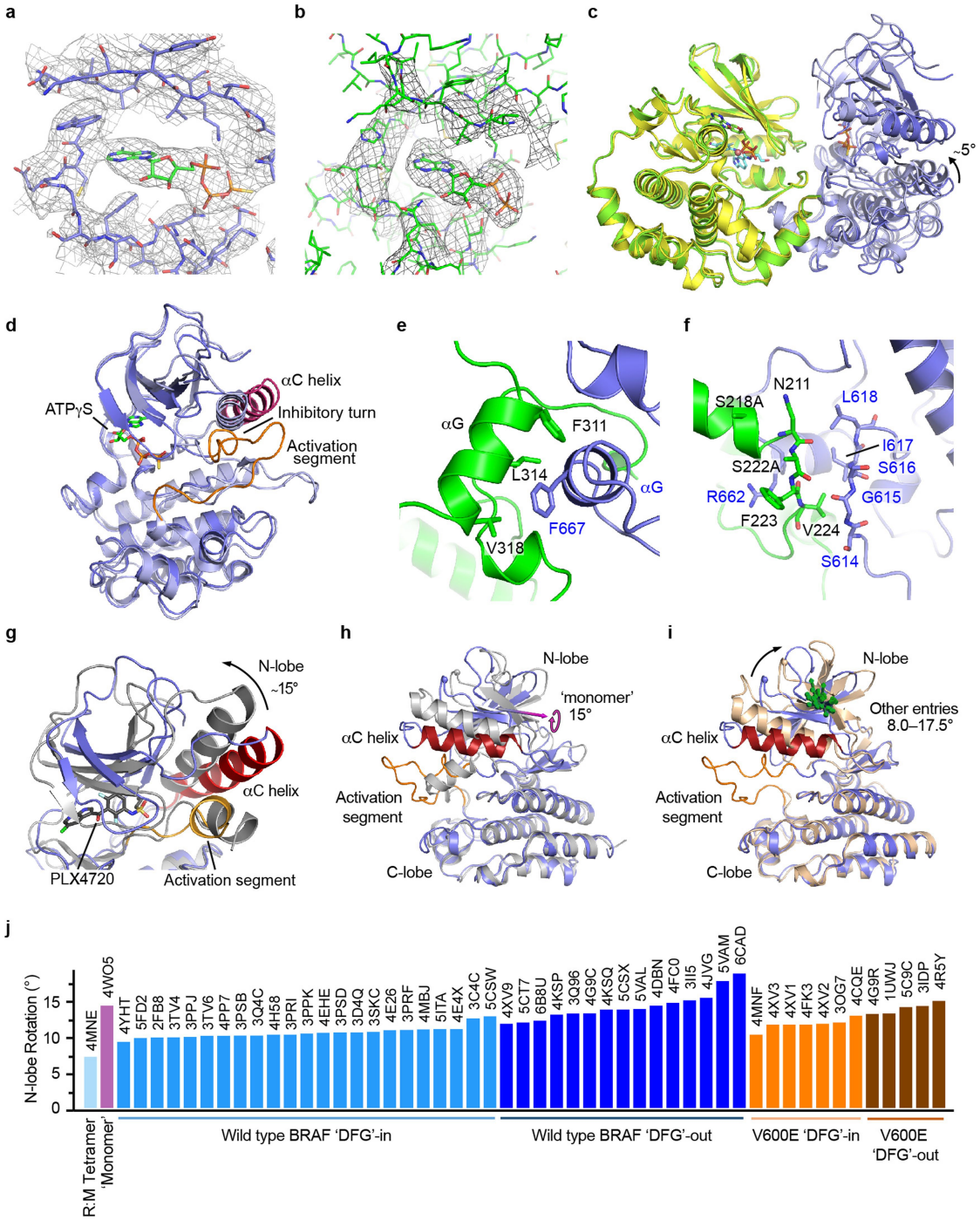
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | 14-3-3 domain sequence alignment and interactions of the CRD domain in the autoinhibited state. a, Sequence alignment of insect-cell (*S. frugiperda*) and human 14-3-3 isoforms. Secondary structure is indicated above the alignment. Identically conserved residues are shaded red. Symbols above the alignment indicate contacts with the BRAF CRD domain (purple squares), kinase domain (blue triangles), and pS365 or pS729 segments (black circles). **b**, Interactions of the CRD domain. Domains that contact the CRD are shown with a transparent surface and the CRD domain is shown as a purple ribbon with grey spheres representing bound zinc atoms. Sidechains are shown for CRD residues that correspond to 7 (of 11 total) residues identified in an alanine scanning mutagenesis study of the CRAF CRD domain[27]. Alanine mutations in the corresponding residues increased RAS(G12V)-dependent activation of CRAF. Two mutations in this study fully activated CRAF in the absence of RAS(G12V); the corresponding BRAF residues are F247 and D249. F247 makes hydrophobic contacts with both the kinase C-lobe and the 14-3-3 domain, whereas D249 is positioned to form a salt bridge with R691 in the kinase C-lobe. The remaining four residues are also at sites of interdomain contacts but are not illustrated (T241, K253, Q262 and K267).



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Interactions of the 14-3-3 dimer with BRAF in the autoinhibited state. a–**d**, Cryo-EM density is shown at key sites of interaction that stabilize the autoinhibited complex, and domains are coloured as in Fig. 1. **a**, A portion of the interface between the CRD and 14-3-3 domain. **b**, The pS365 segment (CR2) bound in the recognition groove of the 14-3-3 domain. **c**, Contact between the α8–α9 loop of the 14-3-3 domain and the BRAF kinase domain. **d**, The C-terminal pS729 segment coordinated in the opposite recognition groove of the 14-3-3 dimer. The map is contoured at the same level in **a**–**d**. **e**, Front and back views of the reconstruction. We observe continuous density connecting the C terminus of the BRAF kinase and the pS729 14-3-3 binding site (inset). **f**, Comparison of the binding mode of the pS365 segment in the present structure with that in a previously determined crystal structure of an isolated CRAF peptide bound to 14-3-3ζ (PDB ID: 3NKX). The corresponding region of the present structure (the pS365 segment is shown with orange carbon atoms and the 14-3-3 domain is shown in tan) is superimposed on the 3NKX crystal structure (shown in blue and cyan), revealing a close correspondence in conformations of the bound peptides. **g**, The BRAF CRD is a hot spot for RASopathy mutations, which map to sites of contact between the CRD (purple), kinase (blue) and 14-3-3 domains (tan), and are expected to destabilize the autoinhibited assembly. Sites of RASopathy mutations are shown in stick form and are labelled. RASopathy mutations in the BRAF kinase domain (Q709) and CR2 region (red, S365) are also expected to destabilize these inhibitory intramolecular contacts.



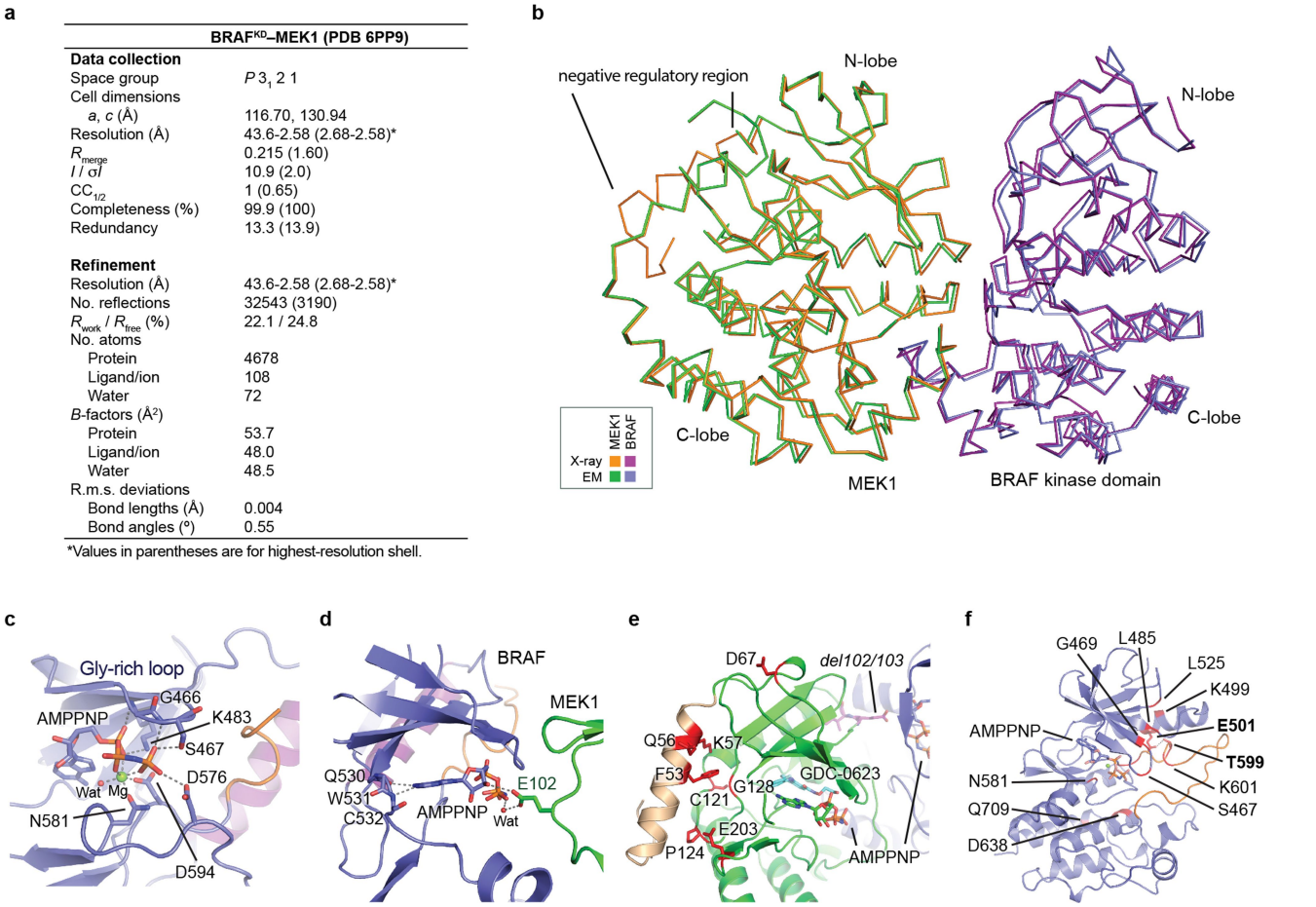
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Additional views and analysis of the BRAF and MEK kinase domains in the autoinhibited BRAF–MEK1–14-3-3 complex. a, Cryo-EM density map in the region of the BRAF active site showing bound ATP-γ-S. **b**, Cryo-EM density map in the region of the MEK1 active site indicating bound ADP, which is probably hydrolysed from ATP-γ-S. Maps in **a** and **b** are contoured at the same level. **c**, Superposition of the BRAF–MEK1 component of the present autoinhibited cryo-EM structure (green and dark blue) with the previously reported crystal structure of a BRAF and MEK1 kinase domain complex (yellow and light blue; PDB ID: 4MNE). The superposition is based on the MEK component of the structures, and it reveals a relative rotation of BRAF of approximately 5° about the C-lobe contact. **d**, Superposition of the BRAF kinase domain from the present structure with that of previously isolated BRAF–MEK kinase complex (PDB ID: 4MNE). Note that the present structure (dark blue, with C-helix coloured purple and the activation segment orange) exhibits key features of an autoinhibited state (C-helix out, with an inhibitory turn in the activation segment), whereas the previous structure (light blue) adopts an overall active conformation. **e**, Detailed view of a portion of the C-lobe contact between BRAF (blue) and MEK1 (green). **f**, Portions of the BRAF (blue) and MEK1 (green) activation segments interact in an anti-parallel orientation. Activating phosphorylation sites in the MEK1 activation loop are substituted with alanine in this structure (S218A/S222A), but neither residue is positioned appropriately for phosphorylation by BRAF. Note that our discussion of these interactions relies in part on the crystal structures referenced to build the atomic model, as the cryo-EM map in this region does not unambiguously define all sidechain conformations. **g**–**j**, Comparison of BRAF kinase domain conformations and relative N- and C-lobe orientations. **g**, Sulfonamide-containing BRAF inhibitors perturb the inactive conformation of BRAF. The BRAF kinase domain in the present structure (blue ribbon, with C-helix coloured red and the activation segment orange) is superimposed on the structure of the BRAF kinase domain crystallized as a monomer with PLX4720 (grey, PDB ID: 4WO5). The superposition is based on the C-lobes of both kinases, revealing an altered orientation of the N-lobe in the inhibitor-bound structure (a rotation of around 15°). Note also that the inhibitory turn in the activation segment helix is replaced by a short helix in the PLX4720 complex. **h**, Alternative view of the superposition shown in **g**, highlighting the axis of rotation (pink arrow) between the N-lobes. **i**, As in **h**, but with a representative inhibitor-bound dimeric BRAF structure superimposed (PDB ID: 5CSW, a dabrafenib complex). The rotation axes for N-lobe rotations of dimer structures are shown as green arrows. Note that the orientation of the rotation axis is similar for all of the dimer structures, but almost orthogonal to that of the monomer structure in **h**. In both **h** and **i**, the Cα atoms of K522 are shown as spheres as a point of reference. **j**, Relative N-lobe rotation of wild-type and BRAF(V600E) crystal structures available in the Protein Data Bank (PDB) are compared with the present nucleotide-bound, autoinhibited structure. As illustrated in **h** and **i**, C-lobes of the BRAF kinase domains were superimposed, and the rotation required to bring the kinase N-lobes into register were calculated using PyMOL. With the exception of 4MNE, all structures compared were determined in complex with inhibitors.

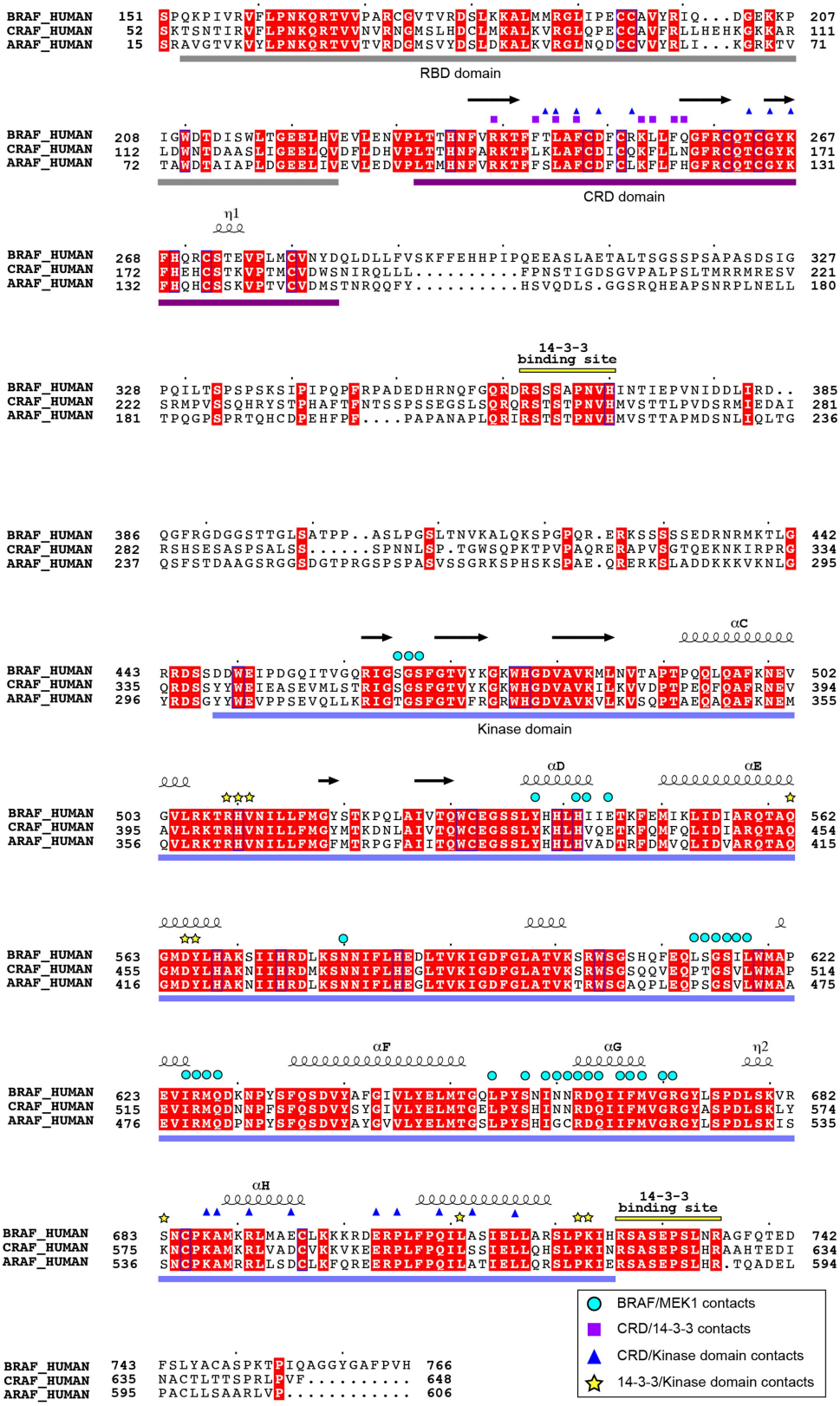
a

BRAF^{KD}–MEK1 (PDB 6PP9)	
Data collection	
Space group	$P\,3_1\,2\,1$
Cell dimensions	
a, *c* (Å)	116.70, 130.94
Resolution (Å)	43.6-2.58 (2.68-2.58)*
R_{merge}	0.215 (1.60)
$I / \sigma I$	10.9 (2.0)
$CC_{1/2}$	1 (0.65)
Completeness (%)	99.9 (100)
Redundancy	13.3 (13.9)
Refinement	
Resolution (Å)	43.6-2.58 (2.68-2.58)*
No. reflections	32543 (3190)
R_{work} / R_{free} (%)	22.1 / 24.8
No. atoms	
Protein	4678
Ligand/ion	108
Water	72
B-factors (Å^2)	
Protein	53.7
Ligand/ion	48.0
Water	48.5
R.m.s. deviations	
Bond lengths (Å)	0.004
Bond angles (°)	0.55

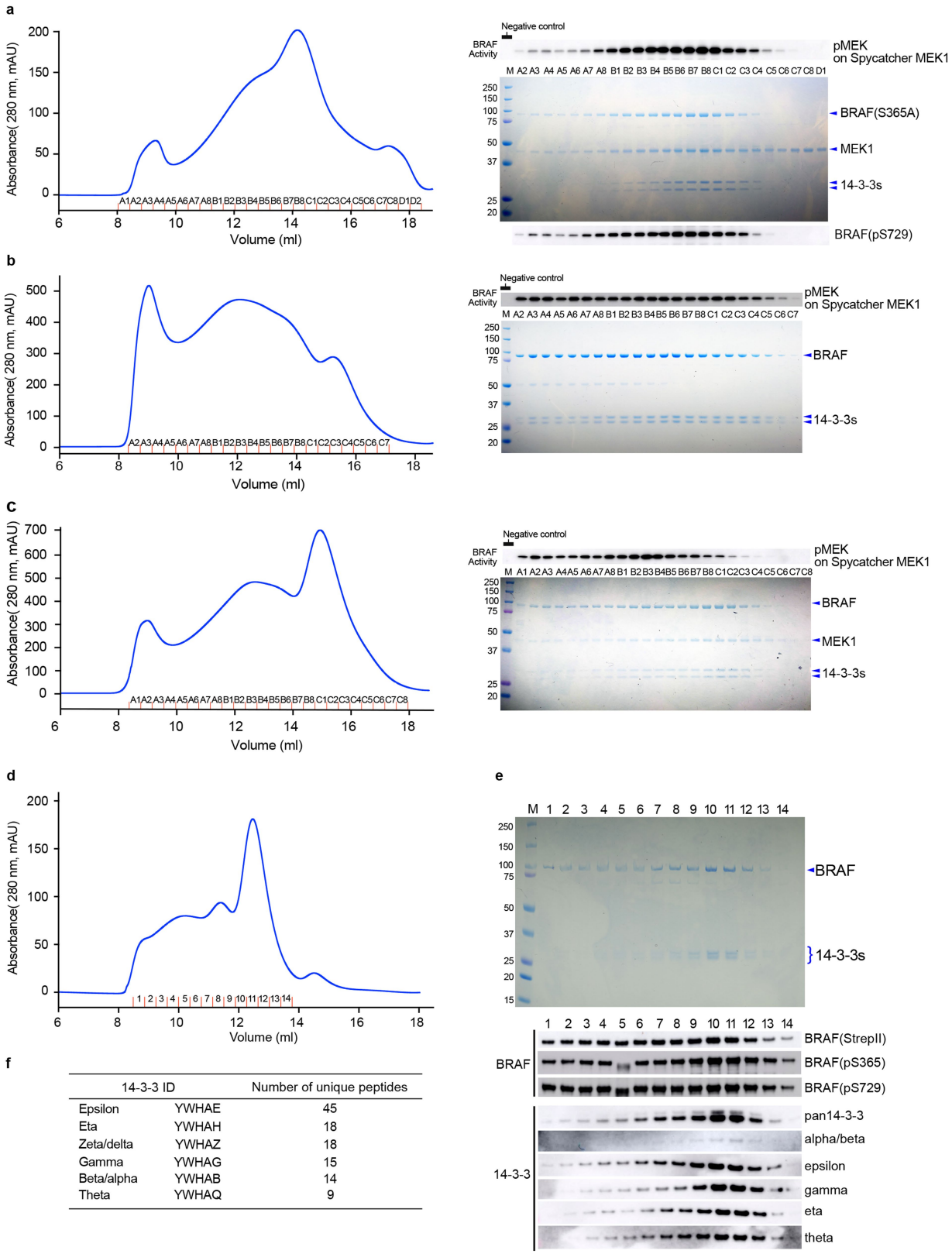
*Values in parentheses are for highest-resolution shell.



Extended Data Fig. 5 | Additional analysis of the crystal structure of the autoinhibited BRAF–MEK1 kinase domain complex. a, Crystallographic data collection and refinement statistics for the structure of the BRAF kinase domain (BRAF^{KD}) in complex with MEK1(S218A/S222A), AMP-PNP and MEK inhibitor GDC-0623. Data were recorded from a single crystal. **b**, The crystal structure of the autoinhibited BRAF–MEK1 kinase domain complex is superimposed on the corresponding region of the autoinhibited cryo-EM structure. **c**, ATP-analogue AMP-PNP is extensively coordinated in the autoinhibited state. Hydrogen bonds from coordinating residues are indicated by dashed lines. **d**, MEK1 residue E102 in the β3–αC loop is positioned to form a hydrogen bond with a ribose hydroxyl of the nucleotide bound in the BRAF active site. **e**, Rare but recurrent oncogenic mutations in MEK1 map to the region of the N-terminal helix. A small, in-frame deletion of two residues (E102, I103) in the β3–αC loop maps to the region of the interface between BRAF and MEK1. **f**, RASopathy mutations in BRAF illustrated in the inactive conformation of the kinase domain. As with oncogenic mutations in many of the same residues, RASopathy-associated mutations will perturb nucleotide binding and/or the stability of the inhibitory turn. Notably, residues E501 and T599 form a hydrogen bond (dashed line) that appears to contribute to the stability of the inhibitory turn.



Extended Data Fig. 6 | RAF sequence alignment. Human ARAF, BRAF and CRAF sequences are aligned and identically conserved residues are shaded red. Secondary structure elements are indicated above the alignment. Symbols above the alignment indicate residues that, in the autoinhibited structure, lie at the interface with MEK1 (cyan circles), the CRD/14-3-3 interface (violet squares), the CRD/kinase domain interface (blue triangles), and the 14-3-3/kinase domain interface (yellow stars).

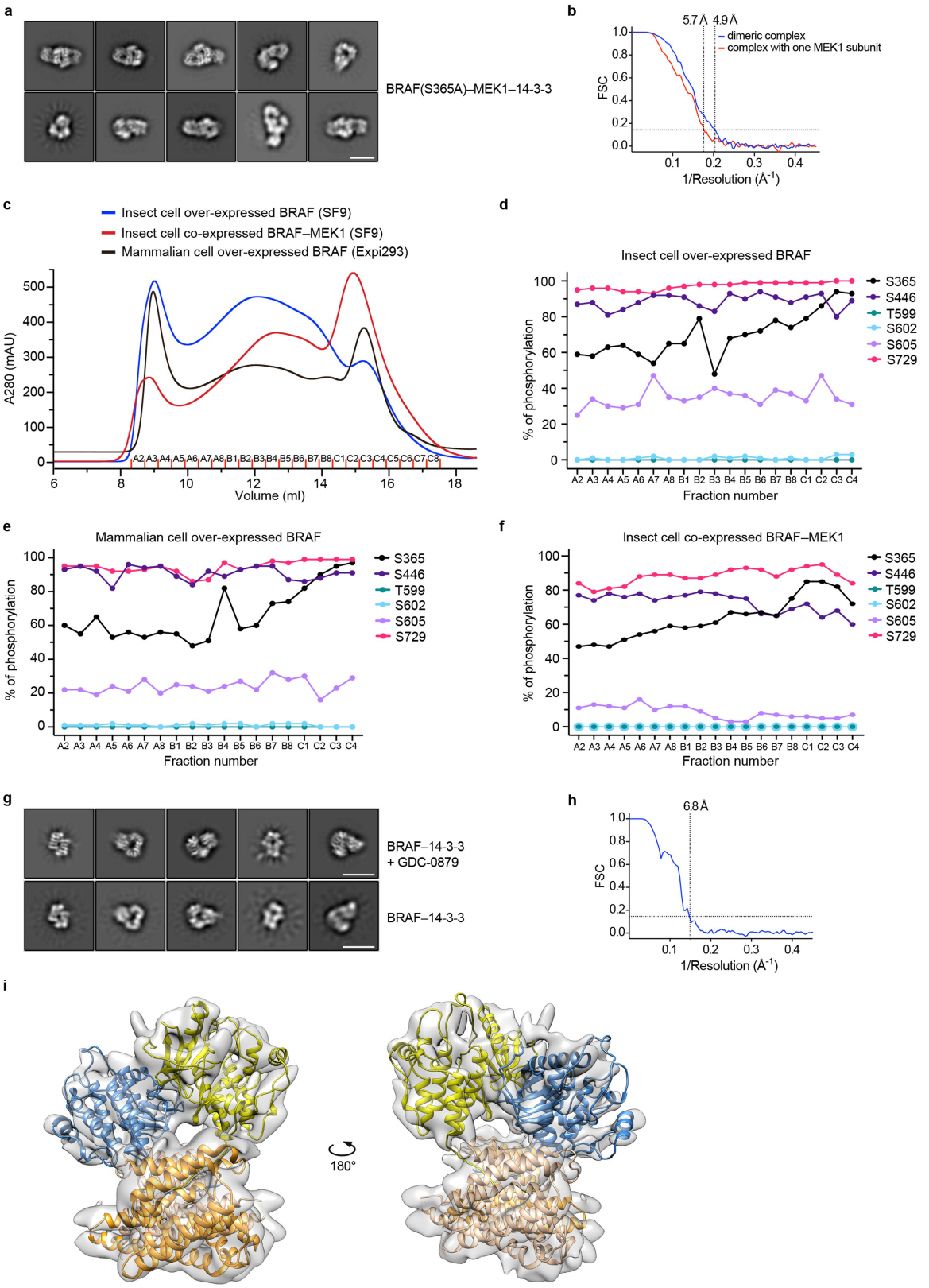


f

14-3-3 ID		Number of unique peptides
Epsilon	YWHAE	45
Eta	YWHAH	18
Zeta/delta	YWHAZ	18
Gamma	YWHAG	15
Beta/alpha	YWHAB	14
Theta	YWHAQ	9

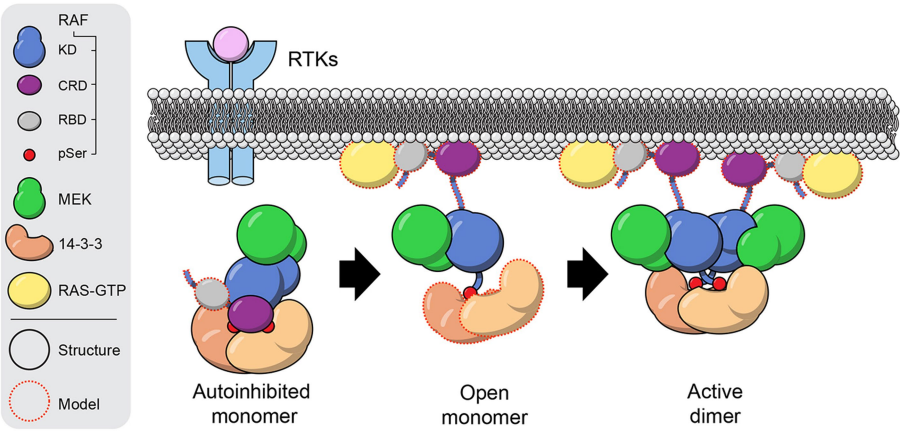
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Purification and characterization of wild-type and BRAF(S365A) complexes. a, BRAF(S365A) was co-expressed with MEK1(S218A/S222A) in insect cells, purified by serial Ni-NTA agarose and StrepTrapHP affinity chromatography, and subjected to SEC on Superose 6 column. The SEC elution trace is shown on the left with a Coomassie-stained SDS–PAGE gel of elution fractions on the right. A parallel gel was blotted with an antibody against pS729 (bottom right). BRAF activity in each fraction was measured in a MEK phosphorylation assay (top right; see Methods for assay details.). **b**, **c**, Side-by-side comparison of wild-type BRAF complexes isolated from insect cells without (**b**) and with (**c**) co-expression of MEK1(S218A/S222A). Complexes were purified by serial Ni-NTA agarose and StrepTrapHP affinity chromatography and subjected to SEC on Superose 6. The SEC elution traces are shown on the left with Coomassie-stained SDS–PAGE gels of elution fractions on the right. BRAF activity in each fraction was measured in a MEK phosphorylation assay as described above (top right). Note that co-expression of MEK1 markedly decreases the void peak and enables the isolation of a late-eluting peak (around 15 ml) with little MEK-phosphorylation activity that corresponds to the autoinhibited BRAF–MEK1–14-3-3 monomer complex (**c**, fractions B8–C3). **d**, Wild-type BRAF was expressed in mammalian HEK293 cells, purified by serial Ni-NTA agarose and StrepTrapHP affinity chromatography, and subjected to SEC on Superdex 200. **e**, Elution fractions from the wild-type BRAF–14-3-3 SEC run in **d** are analysed by SDS–PAGE and western blotting, revealing that BRAF co-purifies with endogenous human 14-3-3 proteins. Fractions were also blotted for total BRAF (anti-StrepII), pS365 and pS729. **f**, Mass spectrometry analysis of trypsin and Lys-C protease digests of peak fractions of the BRAF–14-3-3 complex from HEK293 cells revealed multiple peptide sequences that mapped uniquely to six of the seven human 14-3-3 isoforms. The δ and α isoforms are phosphorylation variants of ζ and β, respectively. For gel source data, see Supplementary Fig. 1. SEC experiments were repeated at least three times (**a**–**e**), activity assays twice (**a**) and once (**b**, **c**), and blotting twice (**e**) with similar results.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Cryo-EM imaging of dimeric, active-state BRAF complexes and mass-spectrometry-based measurement of phosphorylation stoichiometry in BRAF–14-3-3 and BRAF–MEK1–14-3-3 complexes. a, Representative 2D class averages for the BRAF(S365A)–MEK1–14-3-3 complex. Scale bar, 10 nm. **b**, FSC curves for the BRAF(S365A)–MEK1–14-3-3 reconstructions presented in Fig. 4. **c**, SEC (Superose 6) traces for the indicated affinity-isolated BRAF complexes, prepared using insect or mammalian cells as described in Extended Data Fig. 7 and Methods. SEC experiments were repeated at least three times with similar results. **d**–**f**, Per cent phosphorylation of selected BRAF sites in successive elution fractions is plotted for each sample analysed in **c**. Fractional phosphorylation of these sites was measured using a mass-spectrometry-based assay (see Methods). **d**, Wild-type BRAF–14-3-3 complex produced by insect cells. **e**, Wild-type BRAF–14-3-3 complex produced by mammalian cells. **f**, BRAF–MEK1–14-3-3 complex produced in insect cells, prepared by co-expression of wild-type BRAF and MEK1(S218A/S222A). In **d**–**f**, note the high fractional phosphorylation of S729 in all samples, and the negligible phosphorylation of activation segment sites T599 and S602. **g**, Representative 2D-class averages for wild-type BRAF–14-3-3 complexes prepared from mammalian cells with (top) and without (bottom) the addition of BRAF inhibitor GDC-0879 (1 μM). Both samples yielded class averages indicative of the same particle architecture, but those of the drug-treated sample revealed better-defined secondary structure. Scale bar, 10 nm. **h**, FSC curve for the wild-type BRAF–14-3-3 reconstruction. **i**, Single-particle reconstruction of the wild-type BRAF–14-3-3 complex produced in mammalian cells treated with GDC-0879. The reconstruction reveals a back-to-back BRAF kinase domain dimer with a 14-3-3 dimer bridging between its C-terminal pS729 tails. Comparison of these front and back views reveals the highly asymmetric position of the 14-3-3 dimer with respect to the dimerized kinase domains. Imaging experiments in **a** and **g** were repeated twice with independent preparations, and gave similar results.



Extended Data Fig. 9 | Structural snapshots outline a model for RAF activation. The RBD domain is exposed in the context of the autoinhibited BRAF–MEK1–14-3-3 monomer complex, enabling high-affinity binding to farnesylated, GTP-loaded RAS at the plasma membrane. We propose that 'extraction' of the CRD domain upon binding to prenylated RAS at the membrane is a key step in RAF activation. Without the stabilizing interactions of the CRD domain, the 14-3-3 domain can release from the BRAF kinase domain and pS365 segment to form an 'open' monomer. We expect the RAF–MEK kinase module of the open monomer to maintain its inactive, ATP-bound conformation as observed in the crystal structure described here. Finally, the 14-3-3 domain can rearrange to bind the C-terminal pS729 sites of two open RAF molecules, driving formation of the active, back-to-back RAF dimer. As illustrated here, the stoichiometry of 14-3-3 binding changes upon activation, but we do not exclude the possibility that a second 14-3-3 dimer remains associated with the complex, for example by bridging the pS365 segments. KD, RAF kinase domain; red circles (pSer) represent the pS365 and pS729 14-3-3 binding segments.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	BRAF–MEK1–14-3-3	BRAF(S365A)–MEK1–14-3-3		BRAF–14-3-3
		with two MEK	with one MEK	
	(EMD-0541) (PDB 6NYB)	(EMD-20550) (PDB 6Q0J)	(EMD-20552) (PDB 6Q0T)	(EMD-20551) (PDB 6Q0K)
Data collection and processing				
Magnification	130,000 x	36,000 x	36,000 x	105,000 x
Voltage (kV)	300	200	200	300
Electron exposure (e–/Å^2)	~50	~50	~50	~70
Defocus range (μm)	-1.8 – -3.3	-2.0 – -3.0	-2.0 – -3.0	-1.7 – -2.7
Pixel size (Å)	1.06 (2x binned)	1.11	1.11	1.7 (2x binned)
Symmetry imposed	C1	C1	C1	C1
Initial particle images (no.)	3,531,955	2,008,323	2,008,323	365,083
Final particle images (no.)	165,298	425,135	595,672	66,215
Map resolution (Å)	4.1	4.9	5.7	6.8
0.143 FSC threshold				
Refinement				
Initial model used (PDB code)	5FD2, 4MNE, 6PP9, 4FJ3, 1FAR, 3NKX	4MNE, 6NYB	4MNE, 6NYB	4MNE, 3NKX
Map sharpening *B* factor (Å^2)	-225			
Model composition				
Non-hydrogen atoms	8811			
Protein residues	1097			
Ligands	3			
Metals	3			
B factors (Å^2)				
Protein	125.4			
Ligand	146.2			
R.m.s. deviations				
Bond lengths (Å)	0.014			
Bond angles (°)	1.750			
Validation				
MolProbity score	2.18			
Clashscore	6.69			
Poor rotamers (%)	3.66			
Ramachandran plot				
Favored (%)	94.43			
Allowed (%)	5.39			
Disallowed (%)	0.19			
Model vs Data				
CC (mask)	0.74			
CC (box)	0.75			
CC (peaks)	0.62			
CC (volume)	0.73			
Mean CC for ligands	0.77			

nature research

Corresponding author(s): Michael J. Eck, Hyesung Jeon

Last updated by author(s): Sep 17, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	SerialEM for Krios microscope and K2, K3 detector
Data analysis	Standard widely available software was used for structure determination, including RELION version 3, crYOLO, Coot and PHENIX. References are provided in the Methods references.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The cryo-EM MAP and atomic coordinates have been deposited with the EMDB and PDB, respectively (EMD-0541, EMD-20550, EMD-20552, EMD-20551, PDB: 6NYB, 6Q0J, 6Q0T, 6Q0K, 6PP9).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Not applicable
Data exclusions	Not applicable
Replication	Not applicable
Randomization	Not applicable
Blinding	Not applicable

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Antibodies
Vender
Cat#/RRID
Mouse monoclonal Penta-His Antibody
Qiagen
34660
RRID:AB_2619735
Rabbit polyclonal BRAF antibody
ThermoFisher Scientific
PA5-14926
RRID:AB_10975898
Rabbit monoclonal Anti-BRAF (Phospho S729)
abcam
Ab124794
RRID:AB_10976055
Rabbit polyclonal anti Phospho-CRAF(S259) antibody
Cell Signaling Technology
9421S
RRID:AB_330759
Rabbit polyclonal Anti-Strep-tag II
abcam
Ab76949
RRID:AB_1524455
Mouse monoclonal DYKDDDDK tag (9A3)
Cell Signaling Technology
8146S

RRID:AB_10950495
Rabbit polyclonal Anti 14-3-3 Beta/alpha
Cell Signaling Technology
9636S
RRID:AB_560823
Rabbit monoclonal Anti 14-3-3 zeta/delta(D7H5)
Cell Signaling Technology
7413S
RRID:AB_10950820
Rabbit monoclonal Anti 14-3-3 gamma (D15B7)
Cell Signaling Technology
5522S
RRID:AB_10827887
Rabbit polyclonal Anti 14-3-3 epsilon
Cell Signaling Technology
9635S
RRID:AB_2217758
Rabbit monoclonal Anti 14-3-3 eta (D23B7)
Cell Signaling Technology
5521S
RRID:AB_10829034
Rabbit polyclonal Anti 14-3-3 tau
Cell Signaling Technology
9638S
RRID:AB_2218251
Rabbit polyclonal Anti 14-3-3 (pan)
Cell Signaling Technology
CAT#8321S
RRID:AB_10860606
Rabbit polyclonal, Anti-phospho –MEK1/2 (S217/221) antibody
Cell Signaling Technology
Cat#9121S;
RRID: AB_330745
Rabbit polyclonal, Anti- MEK1/2 antibody
Cell Signaling Technology
Cat#9122S ;
RRID:AB_823567
Anti-mouseIgG, HRP-linked secondary antibody

Cell Signaling Technology
Cat#7076s
RRID: AB_330924
ECL Donkey anti-Rabbit IgG, HRP-linked secondary antibody
GE Healthcare

Cat#NA934V; RRID:AB_772191

Validation

Listed RRID for each antibody.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK293 (Expi293F) cells from Thermo-Fisher Scientific
Authentication	Expi293F cells from were obtained directly from Thermo-Fisher Scientific, and were used only for protein production.
Mycoplasma contamination	negative
Commonly misidentified lines (See ICLAC register)	n/a

Author Correction: Exome sequencing of Finnish isolates enhances rare-variant association power

https://doi.org/10.1038/s41586-019-1726-x

Correction to: *Nature* https://doi.org/10.1038/s41586-019-1457-z

Published online 31 July 2019

Adam E. Locke, Karyn Meltz Steinberg, Charleston W. K. Chiang, Susan K. Service, Aki S. Havulinna, Laurel Stell, Matti Pirinen, Haley J. Abel, Colby C. Chiang, Robert S. Fulton, Anne U. Jackson, Chul Joo Kang, Krishna L. Kanchi, Daniel C. Koboldt, David E. Larson, Joanne Nelson, Thomas J. Nicholas, Arto Pietilä, Vasily Ramensky, Debashree Ray, Laura J. Scott, Heather M. Stringham, Jagadish Vangipurapu, Ryan Welch, Pranav Yajnik, Xianyong Yin, Johan G. Eriksson, Mika Ala-Korpela, Marjo-Riitta Järvelin, Minna Männikkö, Hannele Laivuori, FinnGen Project, Susan K. Dutcher, Nathan O. Stitziel, Richard K. Wilson, Ira M. Hall, Chiara Sabatti, Aarno Palotie, Veikko Salomaa, Markku Laakso, Samuli Ripatti, Michael Boehnke & Nelson B. Freimer

In this Article, several errors have been drawn to our attention. In the author list, some of the affiliation numbers were incorrect (see PDF version of the original Article for affiliation numbering): Robert S. Fulton's affiliations should be 2,13 rather than 2; Hannele Laivuori's affiliations should be 8,35,36 rather than 7,35,36; Chiara Sabatti's affiliations should be 10,39 rather than 9,39; Aarno Palotie's affiliations should be 8,40,41 rather than 7,40,41; and Samuli Ripatti's affiliations should be 8,11,41 rather than 7,11,41. In addition, the faculty of affiliation 36 should read 'Faculty of Medicine and Health Technology' rather than 'Faculty of Medicine and Life Sciences'. In the main text on page 326, the protein consequence for the associated *KRT40* variant, listed as 'Ser32Pro', should be 'Ser328Pro'. In the legend for Extended Data Fig. 7, the definitions of SuK and Nfi were missing and have been added. To clarify our description of the Finnish map in Fig. 3, the last line of the legend should read 'Birthplaces of carrier and non-carrier individuals were plotted on a map of Finland, including regions of Finland, later ceded, as they existed before the Second World War.' rather than 'Birthplaces of carrier and non-carrier individuals were plotted on a map of Finland, including regions that were ceded before the Second World War.'. In the 'Apolipoprotein B' section of the Supplementary Information, the protein consequence for the associated *APIM2* variant, listed as '418Asn', should be 'Tyr418Asn'. In the 'Multiplicity adjustment procedure' section of the Supplementary Information, we corrected the annotation in step II. (a) to correctly identify the ordered *P* values (*m*) for all 64 traits at each variant. These errors have all been corrected online.

Author Correction: LKB1 loss links serine metabolism to DNA methylation and tumorigenesis

https://doi.org/10.1038/s41586-019-1696-z

Correction to: *Nature* https://doi.org/10.1038/nature20132

Published online 31 October 2016

Filippos Kottakis, Brandon N. Nicolay, Ahlima Roumane, Rahul Karnik, Hongcang Gu, Julia M. Nagle, Myriam Boukhali, Michele C. Hayward, Yvonne Y. Li, Ting Chen, Marc Liesa, Peter S. Hammerman, Kwok Kin Wong, D. Neil Hayes, Orian S. Shirihai, Nicholas J. Dyson, Wilhelm Haas, Alexander Meissner & Nabeel Bardeesy

In this Article, there were several errors, as follows. The 'Gene Expression Profiling' section of the Supplementary Methods should have included additional information about differential expression analysis, gene set enrichment analysis (GSEA) analysis and the identification of untranslated regions (UTRs). The corrected paragraph is as follows: 'RNA-sequencing was performed using total RNA isolated from two independent cell lines from the K (three replicates total) or KL (four replicates total) genotypes or from the two independent KL lines transduced with full-length *LKB1* cDNA (rescue, two replicates). RNAseq library-preparation and sequencing were performed by the Tufts University Genomics Core Facility. Data were processed using a standard RNA-seq pipeline that used Tophat2[9] to align the reads to mm9, and the Cufflinks suite[10] to calculate expression values and differential expression. Gene Set Enrichment Analysis (GSEA) (http://www.broadinstitute.org/gsea/index.jsp) of the expression data was used to assess enrichment of the KEGG as well as the SGOC geneset[11–13]. There were 2,520 differentially regulated autosomal genes between K and KL samples based on *q*-value as reported by the cufflinks suite (see Supplementary Table 1 of this Amendment). This list was uploaded to the UCSC Genome Browser to extract promoter, intron, exon and UTR sequences. 2,443 of the total 2,520 genes were identified by the algorithm and were associated with 4,706 UTRs. In all cases, pairwise GSEA was performed by creating lists of genes using the FPKM value reported by cufflinks of K to KL or KL to rescue and *P* values were obtained by permuting the gene set (1,000 permutations). To calculate the statistical significance of SGOC pathway enrichment, the SGOC genelist[13] was added to the KEGG signature list and GSEA was performed using this modified KEGG signature list. Raw sequencing files can be found under the Superseries record GSE86145 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE86145).' The Supplementary Information to this Amendment contains the Supplementary Table 1 cited above.

In addition, the following sentence should have been included at the end of the 'Liquid Chromatography Mass Spectrometry' section of the Supplementary Methods: 'Proteomics data were uploaded to https://massive.ucsd.edu/ and can be found under the accession number MSV000082186.'

In the 'Materials' section of the Supplementary Methods, the text 'PSAT1 (sc-133929) from Santa Cruz' should have been added after the information on the 5-hydroxymethycytosine antibody.

The following information should have been provided at the end of the 'SDS-PAGE Analysis' section of the Supplementary Methods: 'For Extended Data Fig. 1k (GLUT1, actin), Extended Data Fig. 2e (LKB1, GLDC, PSAT1, actin), Extended Data Fig. 5c (DNMT1, DNMT3A, actin), Extended Data Fig. 5h (H3K36me3, H3K27me3, H3K4me3, H3), Extended Data Fig. 7f (LKB1, total AMPK, pAMPK (T172), actin), Extended Data Fig. 7g (pACC (S79), total ACC, actin) and Extended Data Fig. 7w (p-p70S6K (T389), p70S6K, actin), samples were derived from each corresponding experiment and blots were processed in parallel. For Extended Data Fig. 7f, the actin membrane was stripped and reprobed for total ACC and the pAMPK membrane was stripped and reprobed for pACC (S79).'

Knockdown efficiencies for shRNAs against AMPKa1 and AMPKa2, and against DNMT1 and DNMT3A (as used in experiments in Extended Data Figs. 7 and 9, respectively) were not provided in the original Article, and are now included as Supplementary Fig. 1 in the Supplementary Information to this Amendment.

In Extended Data Fig. 7f, the blot for AMPKα was inadvertently vertically flipped. See Fig. 1 of this Amendment for a corrected version of the panel. In the top middle (total AMPK) and bottom left (pAMPK (T172)) blots of the panel for Extended Data Fig. 7f in Supplementary Data Fig. 1 of the original Article, the molecular mass markers for total AMPK and pAMPK were mislabelled, and the bottom three markers should have been '25', '37' and '50' kDa instead of '37', '50' and '75' kDa.

In the SGOC network diagram in Extended Data Fig. 4a, the AHCY enzyme should have been shown catalysing the reaction from SAH to HCY instead of the reaction from HCY to Met.

In the legend to Extended Data Figs. 4g and 5a, the sentence: 'The data plotted are expressed as mean-centred values' should have stated: 'The data plotted are expressed using a min-to-max relative colour scheme'.

In the legends to Extended Data Figs. 1k, 2e and 5c, the text: 'Actin was used as the loading control' should have stated: 'For western blot analyses, samples were derived from the same experiment and blots were processed in parallel. Actin was used as the sample processing control.'. Similarly, the legend to Extended Data Fig. 7f, g and w should have included the text: 'Samples were derived from the same experiment and blots were processed in parallel. Actin was used as the sample processing control.'

The legend to Extended Data Fig. 5h should have included the text: 'Samples for H3K36me3, H3K27me3 and H3 were derived from the same experiment. Samples for H3K4me3 and H3 were derived from the same experiment. Blots were processed in parallel and H3 was used as the sample processing control.'.

The Supplementary Information of this Amendment contains Supplementary Table 1 and Supplementary Fig. 1, as described above. The original Article has not been corrected online.

Supplementary information is available in the online version of this Amendment.

Fig. 1 | This is the original, incorrect published Extended Data Fig. 7f, and the corrected Extended Data Fig. 7f. The blot for AMPKα was inadvertently flipped vertically in the published figure.

Work

PhD survey

Graduate students: the tortuous truth
go.nature.com/phdsurvey2019

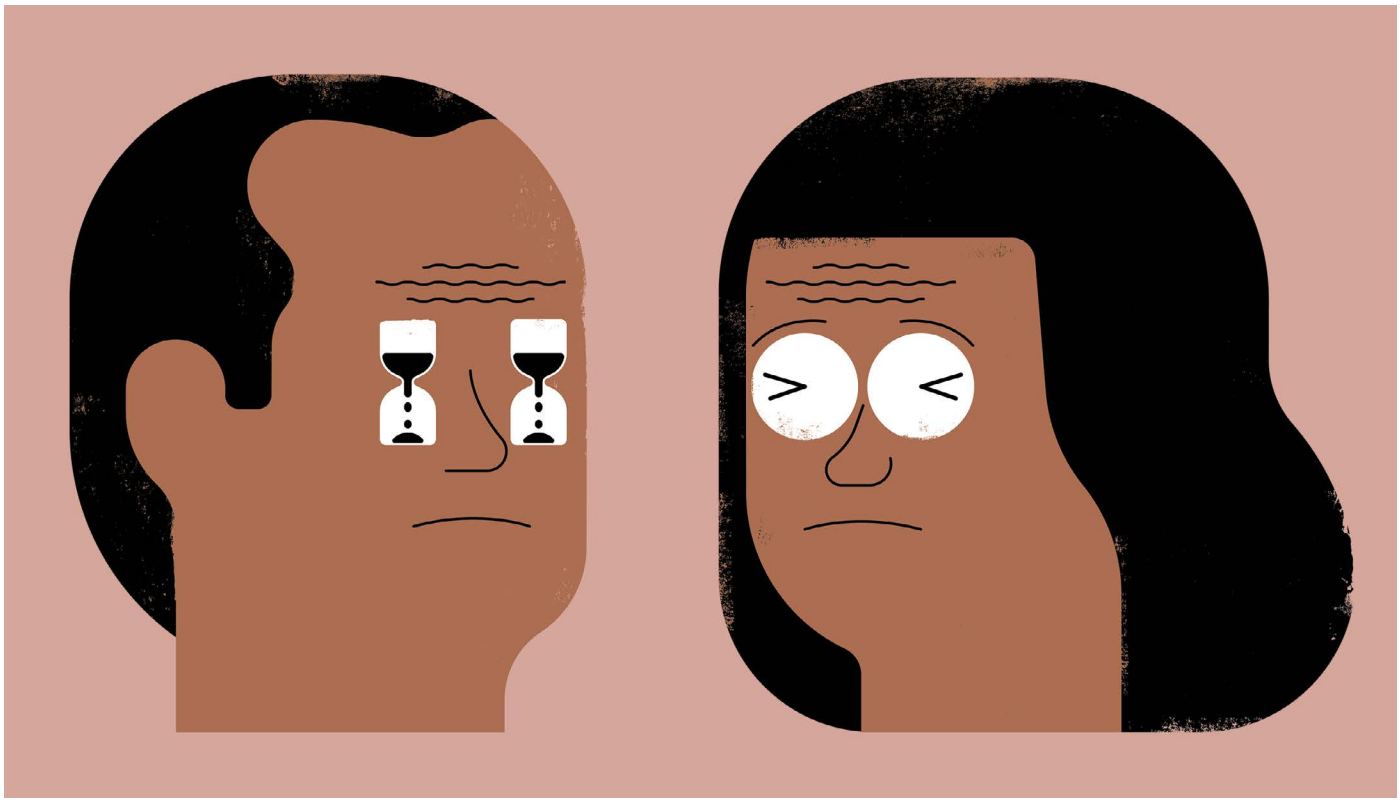


ILLUSTRATION BY THE PROJECT TWINS

JUST A MINUTE... PHD STUDENTS VOICE CONCERNS ON MENTORING

In this second article to mark *Nature*'s 2019 graduate survey, respondents call for more one-to-one support and better career guidance. **By Chris Woolston**

When Peter Butler started his PhD programme in physics at the University of Bristol, UK, he saw himself spending many hours at a whiteboard working on problems, with his supervisor by his side. Those long hours of togetherness never materialized. In that sense, he says, "I didn't get what I expected." However, he adds that his supervisor gave him plenty of good strategic advice and helped him to get published. And having to turn to other people for support was useful, he adds. "I had to act like a scientist."

Butler was one of more than 6,300 graduate students worldwide who responded to *Nature*'s fifth biennial PhD survey. These students had much to say about the state of mentorship at their institutions and in the scientific community. Their answers and free-text comments made clear that they often aren't getting what they expect, or need, from their supervisors. The full data set is available at go.nature.com/2nqjndw. One telling statistic was that nearly one in four said they would change their supervisor if they could start their programme again; the 2017 figure was similar.

The survey — created with Shift Learning, a London-based market-research company — had its bright spots. Overall, 67% of respondents said they were satisfied with their relationship with their supervisors, with 41% of those in Africa and South America saying they were very satisfied. Some are especially grateful. "When I started my PhD, I didn't know about all of the possibilities," says Marina Kovačević, a PhD student in physical chemistry at the University of Novi Sad in Serbia. Now, she hopes to run her own laboratory, a goal that her co-supervisors encourage by letting her help to write proposals and take on other tasks of a lab leader. "She is truly one of the most devoted PhD students," says one supervisor, Branislav Jovic.

But roughly one-fifth of respondents said that they were dissatisfied with their supervisor relationship, a disconnect that threatens their future as well as their present. "Students who are effectively mentored outperform those who aren't," says Ruth Gotian, assistant dean for mentoring at Weill Cornell Medical College in New York City. A coming report from the US National Academies of Sciences, Engineering, and Medicine notes that positive mentorship is the "most important factor in completing a STEM [science, technology, engineering or mathematics] degree". The report also cites studies showing that effectively mentored students are more likely to publish papers,

and more likely to finish their PhD programme.

Luckily for students, mentorship needn't be a one-person job. The survey results underscore the importance of networks that can fill in gaps when a supervisor falls short, says Emma Williams, an author and career coach and founder of EJW Solutions, a scientist-advisory company in Cambridge, UK. "PhD students should be encouraged right from the start to have a variety of mentors," says Williams, who earned her degree in medical physics from the University of Cambridge.

No time for career advice

Many graduate students have discovered that not all mentors can devote much time to the job. In the survey, 49% of students reported spending less than an hour one-to-one with their supervisor each week (see 'Brief encounters'). "That's a shocking figure," Williams says. Although some students can probably thrive on that amount, or on even less, most could benefit from more direct guidance and attention, she says. She speaks from personal experience; her own highly accomplished PhD adviser didn't have the time to build a strong connection. "He called me by the wrong name in the middle of my PhD," she says. "That was a low point."

Job prospects are a persistent worry for PhD students, but they can't always count on their supervisors to show them the way forward. In the survey, just one-third of respondents said that they were satisfied with the career guidance they received from their mentors and others in their PhD programme, down from 40% in the 2017 survey. When asked how they arrived at their current career decision, just 28% credited advice from their supervisor, down from 34% in the survey two years ago.

Notably, 60% of respondents said that they based their career decision on their own research of the topic. Unfortunately, students who try the do-it-yourself approach probably won't be aware of all of their options, Williams says. "They're only going to google the things that they already have in mind," she adds.

Many advisers seem too preoccupied with their own science to offer careers advice, says Nick Valverde, a PhD student in physics at the US National Superconducting Cyclotron Laboratory, located on the campus of Michigan State University in East Lansing. "It's almost impossible to find someone who knows about career trends," he says. "Mentors have a lot on their plates, and trends change." Guidance for careers in industry can be especially hard to come by. Only 28% of respondents said that they had received useful advice for pursuing a career outside academia.

Unready for duty

Part of the problem, Gotian says, is that mentors who have spent their entire careers in academia might not think much about other career paths. "Very often, mentors will try to create 'mini-mes', another version of themselves," she says. But if mentors took off the academic blinkers, they could boost their students' career prospects without much effort, she adds. "They may not have much knowledge of industry, but they probably have contacts that they could connect their students with. That doesn't happen as often as it should."

A further problem is that mentors don't necessarily receive much training in people management, a shortcoming that can contribute to especially dark consequences. In the survey, 21% of respondents reported experiencing discrimination or harassment. The same percentage also reported bullying. Of those, nearly half said that their supervisor was the perpetrator (see 'Bullying from the top'). "In a results-driven culture, you're very dependent on people senior to you to move on with your career," Williams says. "It's very fertile ground for bullying and harassment."

These numbers once again reinforce the need to have more than just one person on a student's side, Williams says. "One of my clients at a prominent university was being bullied," she says. "Finding someone else that she could use as a sounding board really helped her."

With so much at stake, choosing a mentor or a mentorship team can be one of the most important tasks a graduate student can face. Kovačević says that she peppered potential advisers with questions before joining her current lab. "I thought it was my right to ask anything," she says. "And I thought it was their job to answer."

But not all students have that option. "I had no say in choosing my PhD adviser," says Samhita Krishnaswamy, a PhD student in psychology at Jain University in Bengaluru, India. She says that she felt inspired by an accomplished professor in her programme. But he was not her adviser, and she rarely had a chance to speak to him in person. She feels that supervisors, in general, could be better prepared to guide their students. "In India, supervisors need more in-depth skill sets," she says. "They're mostly looking at furthering their own careers. They are very uncomfortable pursuing topics outside of their research area."

Even so, Krishnaswamy says that she's happy with her overall training. She's had the flexibility to study multiple topics in psychology, including the psychology of Indigenous populations of India. "I have everything I need here," she says. "It's given me a foundation to be an independent researcher."

In his third year at the National Superconducting Cyclotron Laboratory, Valverde says that he's building a foundation, too. But it wasn't easy. At first, he was intimidated by the experience and knowledge of Cyclotron researchers. "You're working with someone who has 40 years under their belt," he says. "They're talking about particles and symmetry, and I'm like, man, I know about tension in a rope."

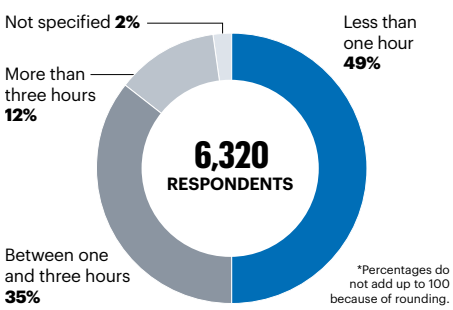
Valverde managed to bridge some of those gaps in his knowledge and form real connections with researchers at the lab — because he had to. Ultimately, he says, science is too challenging to tackle without help. "It could be crippling if you tried to do it alone," he says. "That's where a mentor comes in."

Chris Woolston is a freelance writer in Billings, Montana.

BRIEF ENCOUNTERS

Interactions with a supervisor can be a crucial part of PhD training, but some students get much more individual time than others.

Q: On average, how much one-to-one time do you spend with your supervisor each week?

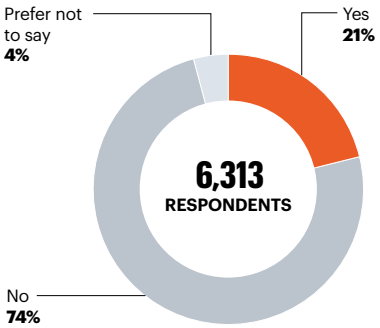


*Percentages do not add up to 100 because of rounding.

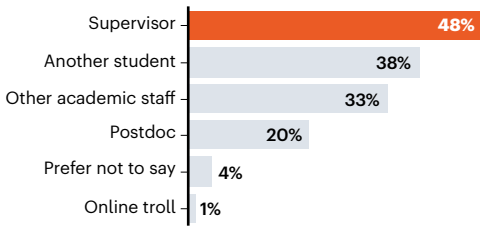
BULLYING FROM THE TOP

A substantial number of PhD students feel bullied in their programmes. Speaking out is difficult — partly because supervisors tend to be the chief culprits.

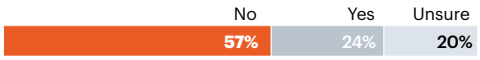
Q: Do you feel that you have experienced bullying in your PhD programme?*



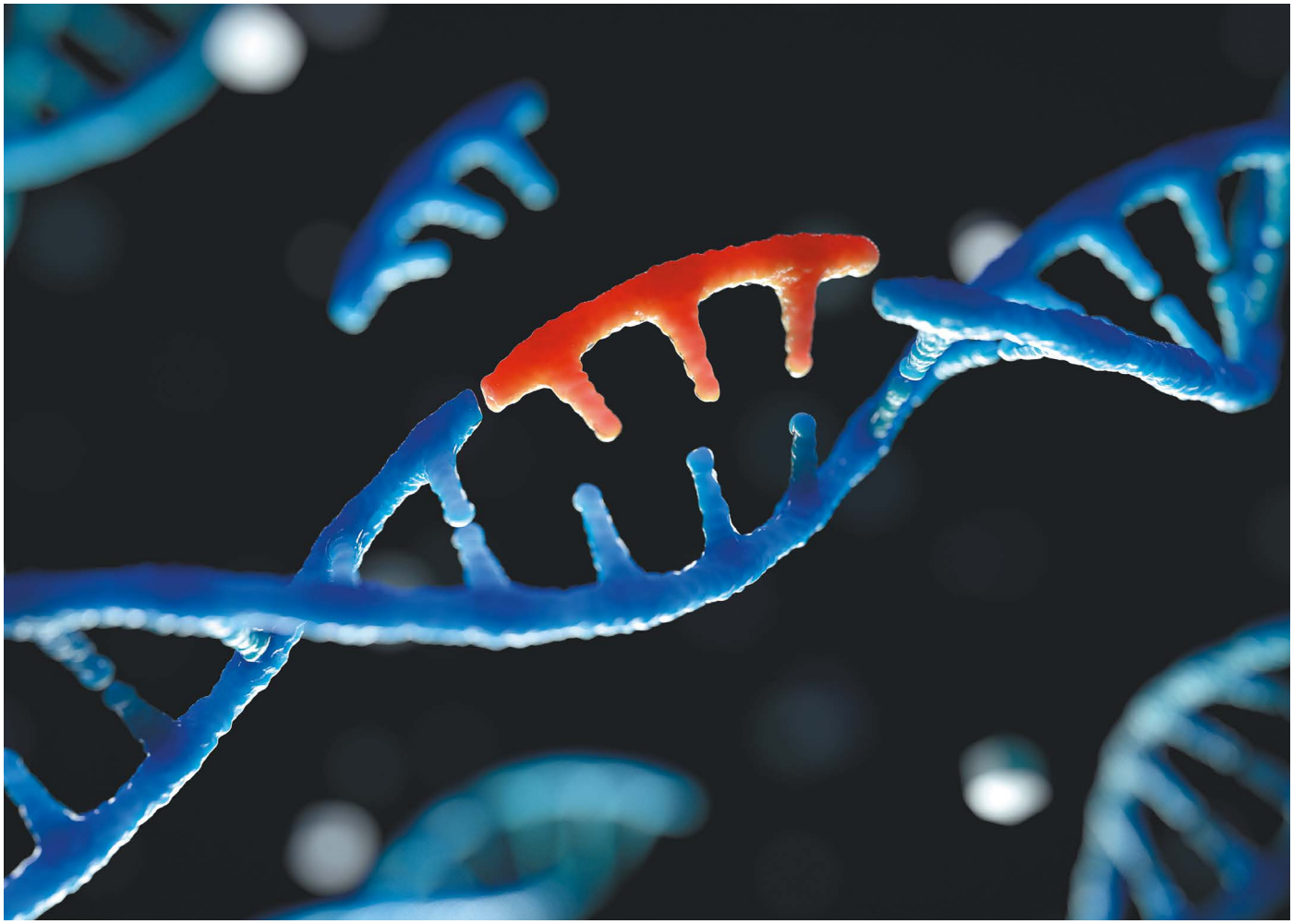
Q: Who was the perpetrator(s)?



Q: Do you feel that you can speak out about your experiences of bullying without repercussions?*



*Percentages do not add up to 100 because of rounding.



GETTY

FIXING GENOME ERRORS ONE BASE AT A TIME

Genetic base editors can efficiently correct point mutations in cell lines, animal models and perhaps the clinic. **By Sandeep Ravindran**

When Xingxu Huang began thinking about correcting disease-causing mutations in the human genome, his attention turned naturally to CRISPR–Cas9. But it quickly became clear that the popular gene-editing tool wasn't ideal for the majority of human disease mutations, which result from errors in single DNA nucleotides known as point mutations. More than 31,000 such mutations in the human genome are known to be associated with human genetic diseases. But CRISPR is not particularly efficient at correcting them.

Then Huang learnt about base editors, a new class of genome-modifying proteins that excel at single-site mutations.

Base editors chemically change one DNA base to another without completely breaking the DNA backbone. The first cytosine base editor (CBE), which chemically converts a cytosine–guanine (C–G) base pair into a thymine–adenine (T–A) base pair at a targeted genomic location, was developed in 2016 by chemical biologists David Liu and Alexis Komor at Harvard University in Cambridge, Massachusetts[1]. Another researcher in Liu's laboratory, Nicole Gaudelli, developed the first adenine base editor (ABE) a year later[2]; it chemically transforms A–T to G–C base pairs.

"Base editing gives very, very good efficiency, about 40–50% efficiency for cell lines," says Huang, a geneticist at ShanghaiTech University in China. "That's very high efficiency compared with traditional genome editing," which is only one-tenth as efficient, he says.

But base editors are not just more efficient than CRISPR–Cas9; they also cause fewer errors. CRISPR–Cas9 acts as molecular scissors that cut both strands of DNA. As the cell repairs the break, random bases can be inserted or deleted (indels), altering the gene sequence. Large chromosomal segments might even be deleted or rearranged. By altering just a specific nucleotide without making double-stranded breaks, base editors cause fewer unwanted mistakes.

Researchers have applied these tools across the evolutionary tree, from bacteria and yeast to rice, wheat, zebrafish, mice, rabbits and monkeys. They have used them to knock out genes, and to create and correct animal models. They have applied them in very early human embryos in the laboratory. And they might one day use base editors to

treat human genetic diseases.

First, however, researchers have to overcome some key hurdles. Like CRISPR–Cas9, base editors sometimes edit sites other than their target. They are limited in which genomic regions they can edit and what base conversions they can perform. And if they are ever to be used in the clinic, researchers will have to get better at delivering them into tissues.

But improved editors are being developed at a rapid rate. "It's really a testament to how fast researchers have made progress in the field that we now have dozens of base editors that offer expanded targeting scope, improved DNA specificity and reduced off-target activity," says Liu. His base-editor constructs have been sent out to more than 1,000 laboratories around the world, he says, and new papers that use these and related tools appear almost weekly.

Building an editor

To create the first base editor, Komor took advantage of a naturally occurring enzyme called APOBEC1. This enzyme, which is part of the cytidine deaminase family, chemically converts C to uracil (U), an analogue of T that occurs in RNA. Komor fused rat APOBEC1 to a catalytically impaired Cas9 nuclease that is unable to create DNA double-strand breaks. When a guide RNA directs the APOBEC1–Cas9 fusion protein to a target site, the deaminase converts C to U. The cell's DNA-repair system then fixes the resulting U–G mismatch by converting it into a U–A base pair, and ultimately to a T–A pair.

Additional refinements improved the protein's efficiency: these included swapping Cas9 for a Cas9 'nickase' that cuts the G-containing strand, thus nudging the cell to replace the G rather than the U when repairing the U–G mismatch. "That extra modification boosted our efficiencies up to levels that we were happy with," says Komor, who is now at the University of California, San Diego. Dubbed BE3, the resulting protein edits cellular DNA with almost a tenfold higher efficiency than CRISPR–Cas9 and with less than 1% indel formation.

The first ABE was tougher to crack. No known naturally occurring enzymes could chemically convert A to G in DNA. "It was a pretty big ask to create an enzyme that didn't exist and have it work very well," Gaudelli says. Luckily for her, Liu's lab had expertise in using microbes to achieve the rapid directed evolution of proteins. Over seven rounds of evolution and protein engineering, Gaudelli gradually coaxed a bacterial enzyme called TadA, which converts A to G in some RNAs, to accept a DNA substrate and work better in mammalian cells, producing an editor called ABE7.10.

Although they can effect only a subset of possible nucleotide changes, such enzymes can already address the majority of disease-causing point mutations in humans, at least in theory. "The adenine base editor, in particular, corrects the most common kind of point mutation in humans," says Liu, referring to G–C to A–T mutations, which account for about half of all known pathogenic single-nucleotide changes. For the moment, however, the technology is for laboratory use only.

Correcting and creating mutations

In initial studies, Liu's team showed that CBEs could correct point mutations associated with Alzheimer's disease and cancer[1] in mouse and human cell lines with an on-target editing efficiency of 35–75% and a 5% indel rate, compared with CRISPR–Cas9's 0.1–0.3% efficiency and 26–40% rate of indel formation. Using ABEs, Liu's team corrected point mutations responsible for a life-threatening blood-cell disorder called hereditary haemochromatosis, as well as sickle-cell anaemia[2].

Researchers have used base editors to create and correct animal models of human diseases, including Duchenne muscular dystrophy[3–5], progeria[3] and age-related macular degeneration (H. Yang, unpublished observations). "With base editors, it's easy to create an animal model and explore pathogenic mutations all over the genome," says Huang, who has generated mouse models of diseases such as androgen insensitivity syndrome and syndactyly, a condition in which multiple fingers or toes are fused together[6]. Huang was even able to combine CBEs and ABEs in the same mouse embryos, resulting in simultaneous A–G and C–T edits, a trick he achieved using editors with different sequence preferences. "We can handle several mutations simultaneously and with very high efficiencies," he says.

"If you are familiar with genome-editing technology, you are ready to do base editing."

Base editors can also be used to produce gene knockouts. The CRISPR–Cas9 system is particularly adept at creating knockouts, thanks to the natural mechanism most commonly used to repair double-strand DNA breaks. That process can add or delete bases at the cut site, causing the gene sequence to be misread and causing protein synthesis to stop prematurely. But CBEs can convert certain codons — the three-base genetic words that define the sequence of amino acids in a protein — to a stop signal directly, an idea that researchers are exploiting to systematically test the effects of knocking out different genes across the genome[7,8]. As base editors progress towards clinical trials, researchers have begun testing them in non-human primates. In unpublished work, Hui Yang, a developmental biologist at the Chinese Academy of Sciences in Shanghai, has applied base editors in mouse and monkey models of eye diseases, such as age-related macular degeneration, as well as Duchenne muscular dystrophy and Parkinson's disease. "Base editors just cause single-strand breaks, not double-strand breaks, so I really think it's more safe than CRISPR," says Yang.

Base editors could also be used to create high-yield or disease-resistant plant varieties, says Caixia Gao, a plant biologist at the Chinese Academy of Sciences in Beijing. "A single nucleotide change can make some rice plants better use nitrogen in the field, for example," she says.

Building a better editor

Although theoretically similar to a genetic search-and-replace tool, base editors are in practice less precise.

The fact that base editing uses Cas9 for sequence targeting means that it can produce off-target changes, just as CRISPR–Cas9 does. But base-editor specificity is complicated further by the deaminases that actually alter the DNA. These enzymes can modify RNA and single-stranded DNA at sites other than the intended target[9–11]. "We don't know if these effects will be clinically relevant or not, but it's wise to try to minimize any unwanted editing," says Liu.

ABEs apparently show no such off-target effects. This is probably because the ABE deaminase binds more weakly to its target than does the CBE deaminase, and so needs Cas9's help for efficient editing, says Liu. Researchers have now developed higher-fidelity CBEs, such as HF-BE3, with weaker target binding, and found that they have correspondingly lower levels of off-target editing[12].

Base editors can also sometimes edit 'bystander' Cs or As that lie within their 'editing window' — the nucleotide region within which the enzyme works efficiently. Researchers have created editors with narrower or broader windows to enhance or reduce such effects. For instance, YE1-BE3 and YEE-BE3 are modified versions of BE3 with narrower activity windows[13], whereas ABE7.9 (ref. 2) and the CBE BE-PLUS[14] have wider ones.

"If we think about genetic disease correction, we need to have very specific editing, where we need to have this activity window be very narrow, down to one nucleotide," says Gao. But an expanded editing window could be useful for accessing multiple target sites, for instance to introduce several point mutations at once.

Base editors are also relatively limited in terms of the genomic sites that they can target; they can only act near a protospacer adjacent motif (PAM), the short DNA sequence required for successful binding of Cas9 to a DNA target. Because of that restriction, "I believe only about 25% of the pathogenic mutations

in the human genome can be precisely edited or corrected using current tools", says Huang. Researchers have expanded base editors' scope by using directed evolution to create Cas9 proteins that recognize a broader range of PAMs, and by fusing base editors to Cas9 variants with wider PAM compatibility.

And then there is the issue of the limited range of base changes that editors can currently produce. To correct as many genetic diseases as possible, base editors will need to perform additional conversions, such as C to A, C to G, A to C and A to T. Jin-Soo Kim, a biochemist at the Institute for Basic Science in Daejeon, demonstrated this year that ABEs can achieve C-to-G conversion as well as C-to-T and A-to-G conversions in a human kidney cell line[15]. "These results give us a hint on how to make other types of base editors," he says.

Alternatively, researchers could use a new class of genome editors from Liu's lab, called prime editors, which can change any DNA base into any other[16]. Prime editors use a special guide RNA template and Cas9 nickase to direct a reverse transcriptase enzyme to a target site. There, the enzyme makes a new DNA strand from the RNA template and inserts it at the target (see 'Prime corrective'). But there are a lot of unknowns with these tools, "including whether we can successfully do prime editing in animals and whether it will be as generalizable for many different types of cells as base editing", says Liu.

With all these different options, researchers will need to consider their needs carefully to find the best fit for their project. For efficiently disrupting genes or inserting or replacing large DNA sequences, CRISPR–Cas9 is the best bet, says Liu. It has been well studied, has lots of variants with greater specificity or particular PAM affinities, and is already being tested in clinical trials. Prime editors offer the greatest flexibility for creating DNA insertions, deletions, point mutations or combinations thereof. And base editors are ideal for correcting point mutations, providing higher efficiency and causing fewer indels.

"I think all three of these classes of genome-editing agents really have complementary strengths and weaknesses," says Liu. He likens CRISPR–Cas9 to scissors, base editors to pencils, and prime editors to word processors. "I think they all have their own roles in research and in applications such as agriculture and human therapeutics, just as scissors, pencils and word processors all have their own useful and unique roles."

As easy as CRISPR

And just like scissors, pencils and word processors, base editing has been rapidly adopted by the scientific community, a testament to its low barrier to entry. "If you are familiar with genome-editing technology, I think you are ready to do base editing," says Kim.

Researchers can order base editors from the non-profit plasmid repository Addgene. Liu recommends starting with some of the newest editors, such as his lab's BE4Max and ABEMax, which target C and A, respectively. But many others could also fit the bill, he adds, depending on the circumstances. (See Table 1 in ref. 17 for a good starting point.)

Consider PAM specificity and the editing window required to access the target, Liu says. Consider also how much to prioritize reduced bystander editing or off-target effects. Specialized computational tools such as beditor can help researchers to design guide RNAs for their particular target.

Still, base editors don't always work as expected. "Sometimes we have to test a couple of different editors before we find one that likes our target," says Komor. If nothing works, researchers can cut and paste from different base editors to make a custom editor, a process that Komor says is relatively straightforward. "Don't be afraid to make your own."

Whatever the editor, delivering them to cells requires standard genetic techniques, such as transfection, micro-injection and electroporation. "You can deliver them as protein–RNA complexes, as mRNA or as DNA," says Liu. Therapeutic applications, however, will require a different approach.

Conventional viral delivery vectors, such as adeno-associated virus (AAV), carry only limited genetic cargo, and base editors are typically too large to fit. "Our current work is aimed at decreasing the size of the Cas9 and base editor, which I think will broaden its application," says Yang. Alternatively, researchers can split base editors across two vectors, as Kim did to target a mutation in the Duchenne muscular dystrophy gene in adult mice. "We were able to correct the mutation in skeletal muscle," he says[5].

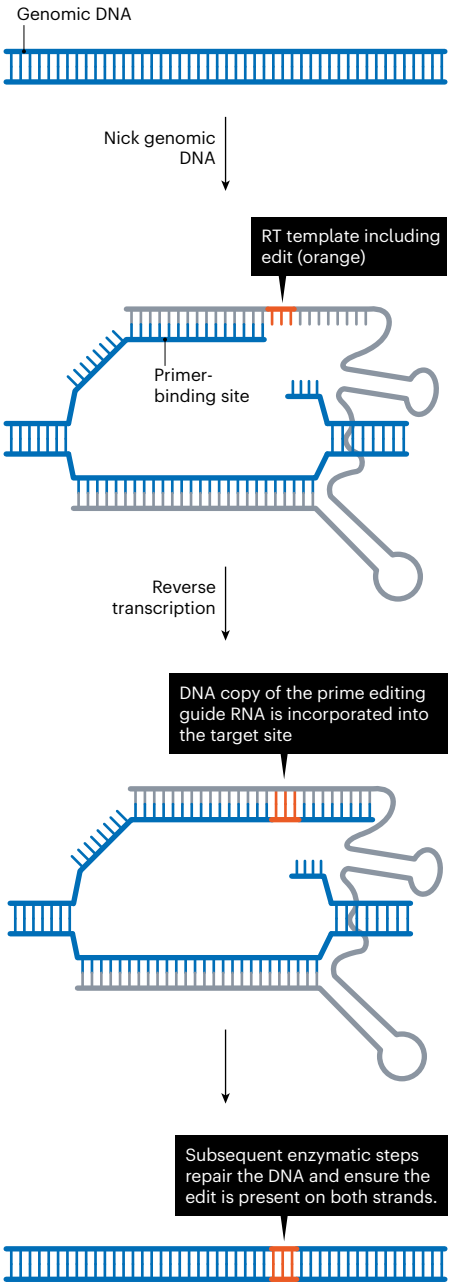
It is early days, but base editors have already become a promising addition to the genome-editing toolset. And they might have more tricks up their sleeves. Some editors, for instance, can act on RNA rather than DNA, opening up the possibility of knocking down or editing mRNA transcripts containing pathogenic mutations. Base editors might also be able to target mutations in mitochondria, which lack the DNA-repair pathways that conventional genome editing relies on, says Kim.

For Gaudelli, such opportunities represent the realization of a lifelong dream. "My motivation for being in the sciences was to make a difference in the world," she says. "I never thought it would be through base editing."

Sandeep Ravindran is a science writer based in Washington DC.

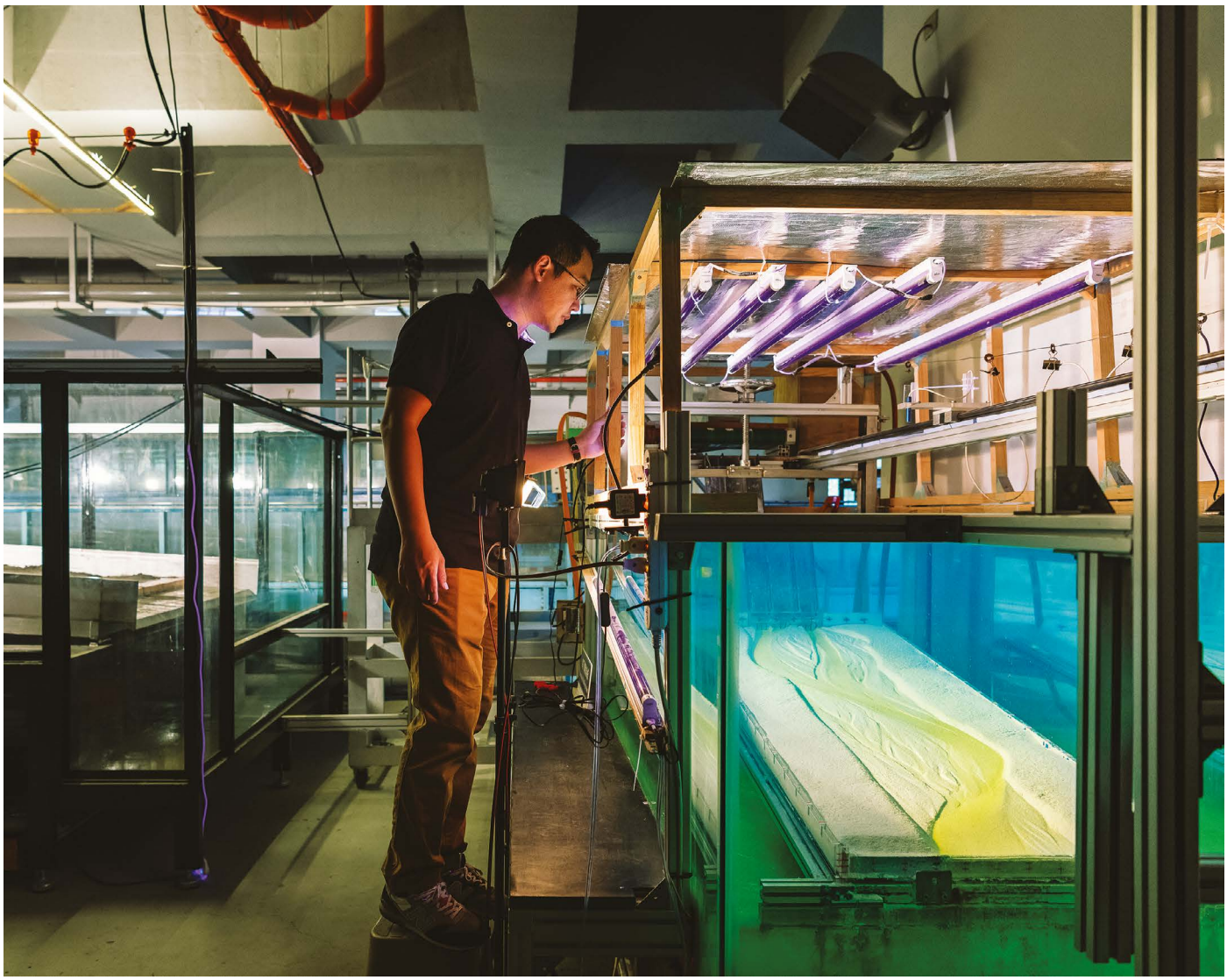
PRIME CORRECTIVE

David Liu's prime editing strategy uses an RNA template and the enzyme reverse transcriptase (RT) to write genomic changes into the DNA.



ANDREW ANZALONE AND DAVID LIU

1. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. *Nature* **533**, 420–424 (2016).
2. Gaudelli, N. M. *et al. Nature* **551**, 464–471 (2017).
3. Liu, Z. *et al. Nature Commun.* **9**, 2338 (2018).
4. Liu, Z. *et al. Nature Commun.* **9**, 2717 (2018).
5. Kim, K. *et al. Nature Biotechnol.* **35**, 435–437 (2017).
6. Ryu, S.-M. *et al. Nature Biotechnol.* **36**, 536–539 (2018).
7. Kuscu, C. *et al. Nature Methods* **14**, 710–712 (2017).
8. Billon, P. *et al. Mol. Cell* **67**, 1068–1079 (2017).
9. Grünewald, J. *et al. Nature* **569**, 433–437 (2019).
10. Zuo, E. *et al. Science* **364**, 289–292 (2019).
11. Jin, S. *et al. Science* **364**, 292–295 (2019).
12. Rees, H. A. *et al. Nature Commun.* **8**, 15790 (2017).
13. Kim, Y. B. *et al. Nature Biotechnol.* **35**, 371–376 (2017).
14. Jiang, W. *et al. Cell Res.* **28**, 855–861 (2018).
15. Kim, H. S., Jeong, Y. K., Hur, J. K., Kim, J.-S. & Bae, S. *Nature Biotechnol.* **37**, 1145–1148 (2019).
16. Anzalone, A. V. *et al. Nature* https://doi.org/10.1038/s41586-019-1711-4 (2019).
17. Rees, H. A. & Liu, D. R. *Nature Rev. Genet.* **19**, 770–788 (2018).



Where I work
Steven Yueh Jen Lai

In my lab, we study the evolution of underwater landscapes. We look at density currents that plunge into an ocean or reservoir, and how sediment and water interact to shape the evolution of deltas, channels and canyons. These kinds of powerful currents occur in extreme events — such as floods or typhoons — and often are too hard to measure from a ship. But we can easily reproduce them here in my lab, using sand and coloured-water flows in a tank.

This is a dream workspace for me. But things were very different in 2014, when my master's student and I were trying to recreate what is essentially a braided river channel on the sea floor. I borrowed space in an old fluid-mechanics lab, and we built a new water tank in a very small corner of this cramped, dark lab.

It was challenging. My student redesigned the small flow boxes that direct the water and tested them again and again. One day, he called me in: "Do you think these look like submarine braided channels?" I said, "My god! You really did it." But even so, working in that old space felt like the end of the world.

We spent a year and a half rebuilding the lab. Now, it is ideal for a flow and sediment experimentalist — wonderful, open and bright. We have enough tools to make an idea become a prototype, and tanks and flumes to test an idea and then to redesign if needed. It's a very positive cycle for me and for my students.

These days, we use very fine-grained sand in a tank to build the continental slope — where the continental shelf dips into the sea — then we inject denser salt water to flow across the slope. We use fluorescent dyes to visualize the water flow and see how it controls erosion or deposition.

During an experiment, this means that the lab is like a darkroom lit up by fluorescent water. It's a wonderful, vivid experience when you see the phenomenon that you generated unfold. In many ways, it's like being on a movie set.

Steven Yueh Jen Lai is an associate professor in the department of hydraulic and ocean engineering at National Cheng Kung University in Taiwan. **Interview by Kendall Powell.**

Photographed for *Nature* by Joe Russo

nature
index
Collaboration and big science

BATTLE LINES

Research rides out geopolitical storms

Paper chasers
Managing the mass-author enterprise

Tracking links
An overview of the strongest partnerships

The tables
Top players in hot fields

nature
index

Collaboration and big science

Nature Index 2019 Collaboration and big science
Nature Index 2019 Collaboration and big science, a supplement to *Nature*, is produced by Nature Research, the flagship science portfolio of Springer Nature. This publication is based on data from the Nature Index, a Nature Research database, with website maintained and made freely available at natureindex.com.

Nature editorial offices
The Campus, 4 Crinan Street,
London N1 9XW, UK
Tel: +44 (0)20 7833 4000
Fax: +44 (0)20 7843 4596/7

Customer services
To advertise with the Nature Index, please visit natureindex.com or email clientservicesfeedback@nature.com

Political instability and escalating trade tensions have dominated 2019. As US–China relations deteriorate and the uncertain consequences of the United Kingdom's planned departure from the European Union loom, science cannot escape the fray. In the United Kingdom, cross-border funding programmes with Europe are increasingly vulnerable, and in the United States, a government-led crackdown to root out alleged espionage is complicating joint efforts between American and Chinese researchers.

The full extent of any damage to research output is not yet clear, but early signs in UK and EU collaborative science are worrying. Between 2015 and 2018, the United Kingdom's annual share of EU research funding fell by €430 million (US$473 million), largely owing to a reduction in grant applications. The number of UK–EU collaborative articles published in journals tracked by the Nature Index grew by just 1.3% from 2016 to 2018, although the stagnation seems unlikely to have been caused by the Brexit vote given the long lead times for research publication. By contrast, US–China collaborative articles jumped by 32% during the same period.

"There shouldn't be any barriers," says US-based neuroscientist, Hongkui Zeng, who relies on imaging expertise in China to investigate brain functioning. The stories of international partnerships you'll read about in this supplement are illustrative of resilience against political pressure. They are stories of mutual admiration, not animosity, and a desire for openness, not secrecy. And there is a lot at stake. In the journals covered by the Nature Index, the number of internationally collaborative articles across the four broad subject areas has risen by between 21% (physical sciences) and 48% (chemistry) since 2012. Publications with thousands of authors are becoming more common in fields known for conducting big science projects, such as high-energy physics, genetics and oncology, as highlighted in this supplement.

This rise in collaborative research has been driven by necessity. As global challenges become increasingly complex, the most valuable teams are those with interdisciplinary skill sets and diverse perspectives. For research, the truism that great things are seldom done alone, has never been more accurate.

Bec Crew
Senior editor

On the cover
Researchers overcome geopolitical obstacles to maintain connections. Illustration by Roberto Cigna

Contents

QIAN ZHAO FOR HARVARD-CHINA PROJECT/ HARVARD GLOBAL INSTITUTE/ TSINGHUA UNIVERSITY

Shared endeavours, such as this energy study by the Havard China Project, continue to flourish.

Science weathers political ill wind

Despite government tensions, research collaboration between China and the United States remains strong. **By Chris Woolston**

Some projects are too big for a single lab. Or, for that matter, a single country. Hongkui Zeng, a neuroscientist at the Allen Institute for Brain Science in Seattle, Washington, is working on an ambitious project that spans the Pacific. Her team is attempting to untangle the subtle structural differences among groups of neurons in the mouse neocortex, where higher cognitive functioning such as sensory perception and spatial reasoning is processed. The pursuit requires major assistance from scientists in China, whose international research presence is strong, despite growing mistrust from the US government.

To get a clearer picture of the mouse neocortex, Zeng depends on high-resolution neuron images from Huazhong University of Science and Technology, in China (HZAU). "They have a unique bioimaging centre," she says. Access to these images allows researchers to more closely examine the shape and structure of the axons that carry electrical messages from neurons. "The morphology helps us understand how neurons form networks and communicate with each other," she says. "It's critical information."

In science as in cortexes, networks are crucial. According to the Nature Index, connections between the United States and China are stronger than ever. The number of papers co-authored by the two countries in the 82 high-quality journals tracked by the index leapt from 3,413 in 2015 to 4,631 in 2018. Chinese authors collaborated with researchers in the United States more often than in any other region, and China is second only to the European Union as the collaborator of choice for researchers in the United States.

These partnerships have formed against a backdrop of political tensions, economic tariffs, and even fears of academic espionage. In September, US federal prosecutors announced the arrest of Zhongsan Liu, head of the New York office of the China Association for International Exchange of Personnel, for his alleged involvement in a conspiracy to fraudulently obtain US visas for Chinese government employees. A month earlier, the Federal Bureau of Investigation arrested Chinese chemist, Feng Tao, an associate professor at the University of Kansas, for allegedly failing to disclose his full-time employment with Fuzhou University in China.

Earlier this year, the MD Anderson Cancer Center in Houston, Texas, sacked three Chinese scientists over alleged theft of research data for China, and *Science* magazine reported that the US National Institutes of Health (NIH) sent letters to 77 institutions warning of collaboration with scientists who may have ties to foreign governments, including China.

Such incidents have caused consternation and unease among Chinese researchers and their collaborators in the United States, Zeng says. But scientific cooperation remains strong for a fundamental reason: each country has intellectual and material scientific resources that transcend political boundaries. "Science needs a free exchange of information," says Zeng. "There shouldn't be any barriers."

Cooperation flows in both directions. In 2018, Zeng, and Allen Institute geneticist, Linda Madisen, were co-authors on a paper that presented a three-dimensional atlas of the cholinergic system in the mouse brain, an important model for Alzheimer's disease research. The study, published in the *Proceedings of the National Academy of Sciences*, was led by researchers at HZAU, and was based on mouse cell lines created by Zeng and Madisen in the United States. "We develop tools and

resources and share them with the scientific community," says Zeng. "They should be available to anybody."

Despite its high-profile warnings regarding foreign research partnerships, collaboration also continues at the NIH. Christopher Buck, a virologist at the National Cancer Institute's Center for Cancer Research, a division of the NIH, credits Chinese researchers for playing a major role in his latest finding: the discovery of 12 new types of small DNA tumour viruses, infectious agents with the potential to cause cancer. He found the first hints of a new virus in fish DNA sequenced by researchers at the BGI Academy of Marine Sciences in Shenzhen, China, led by Chao Bian.

Available to anybody

Bian's team is part of BGI Marine, the agricultural arm of China's BGI Group, a gene-sequencing company that runs the largest genetics research centre in the world. The BGI researchers didn't have viruses in their sights when they originally shared DNA sequences with the NIH researchers. They were investigating the genetic make-up of the green Asian arowana (*Scleropages formosus*), a prized aquarium fish that can fetch several hundred thousand dollars on the market.

When Buck informed them of the intriguing viral sequences, they returned to their samples and sequenced the entire viral genome. "They kindly and bravely trusted that this strange person from across the ocean had a valid hypothesis," he says. A preprint of the study was published in bioRxiv in August.

BGI's sequencing powers fuel a significant portion of China–US scientific collaboration. In the Nature Index, BGI is involved in 7% (see 'Big picture science', S28) of collaborative articles involving at least one corporate institution publishing in genetics between 2015 and 2018.

"I have great admiration for BGI," Buck says. "They're sequencing all sorts of interesting things, and they're sharing what they find. It takes teams of specialists to look at the data and see what's going on." He adds that no one at the NIH has ever warned him about sharing sensitive materials with researchers from China or any other country. "I have no secrets," says Buck. "Taxpayers are paying me to create scientific knowledge and broadcast it as widely as possible."

"They kindly and bravely trusted that this strange person from across the ocean had a valid hypothesis."

Harvard University's long history of scholarship on China and deep ties with the country has set the stage for ongoing cooperation in many scientific fields, says Michael McElroy, a climate scientist and chair of the Harvard China Project, a joint effort of several Harvard colleges established in 1993 to understand and tackle energy, environmental and economic issues in China and beyond.

For the past few decades, hundreds of Chinese graduate students who came to Harvard as part of the project have returned to their home country to become government officials or scientists. Those students are part of a larger trend: nearly 10,000 researchers, mostly of Chinese origin, moved from the United States to China in 2017 alone.

Great admiration

Those networks have led to many collaborations, including a *Nature Sustainability* paper published in July suggesting that China may be five to ten years ahead of schedule in meeting its Paris Agreement pledge to curb carbon dioxide emissions by 2030. The paper was co-authored by McElroy and a group of Chinese researchers led by Haikun Wang, a researcher at Nanjing University and one of McElroy's former graduate students.

McElroy says that much of the funding for the Harvard China Project now comes from the Harvard Global Institute, a sign of change. "In the early days, we had funding from the National Science Foundation and NASA," he says. "Today, there's no possibility of getting funding from official US sources for China-related work. It's discouraged in Washington."

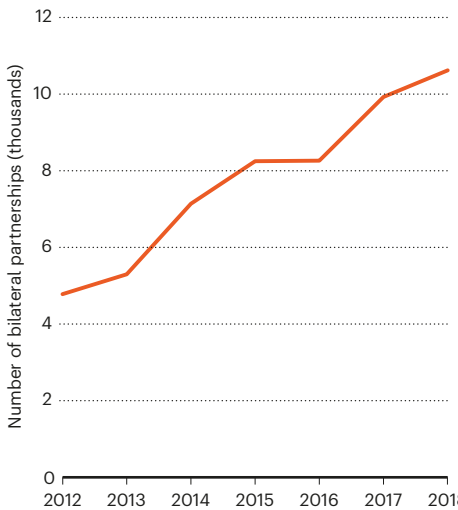
Despite the sometimes uneasy climate, McElroy predicts that collaboration between US and Chinese scientists will continue to propel science in both countries. "The strength of cooperative scholarship has not declined," he says. "If anything, it's growing."

Chris Woolston is a freelance science writer in Billings, Montana.

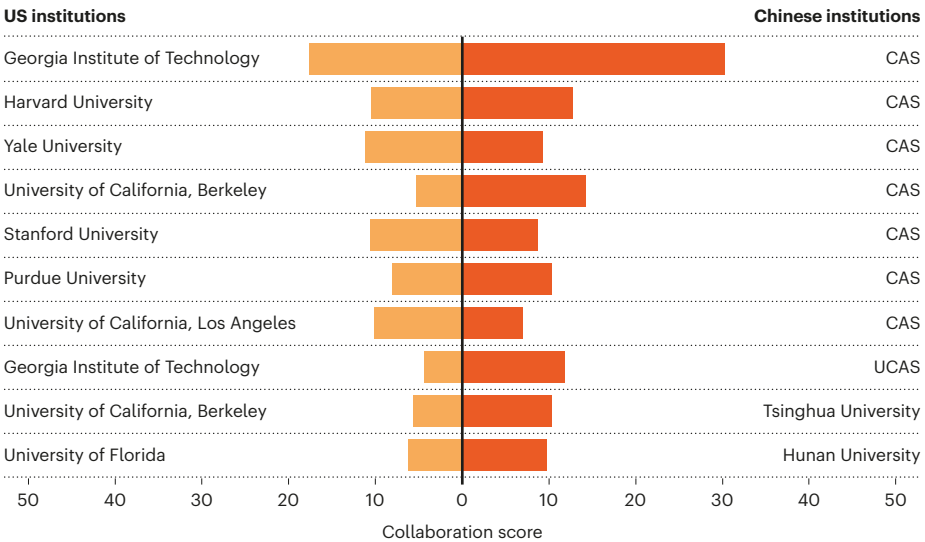
PROLIFIC PAIRING

Growth in the number and strength of institution-to-institution research relationships shows no sign of slowing, despite political tensions between the US and China.

Bilateral partnerships
Research relationships between US and Chinese institutions grew strongly in the 6 years to 2018.



Top cooperators
Among the most productive US-China partnerships in high-quality research in 2018, 7 of the top 10 involve the Chinese Academy of Sciences (CAS).



SOURCE: NATURE INDEX

Big picture science

Across continents and research fields, big science is a global enterprise, yet even for leading collaborators, the strongest partnerships are mainly local.

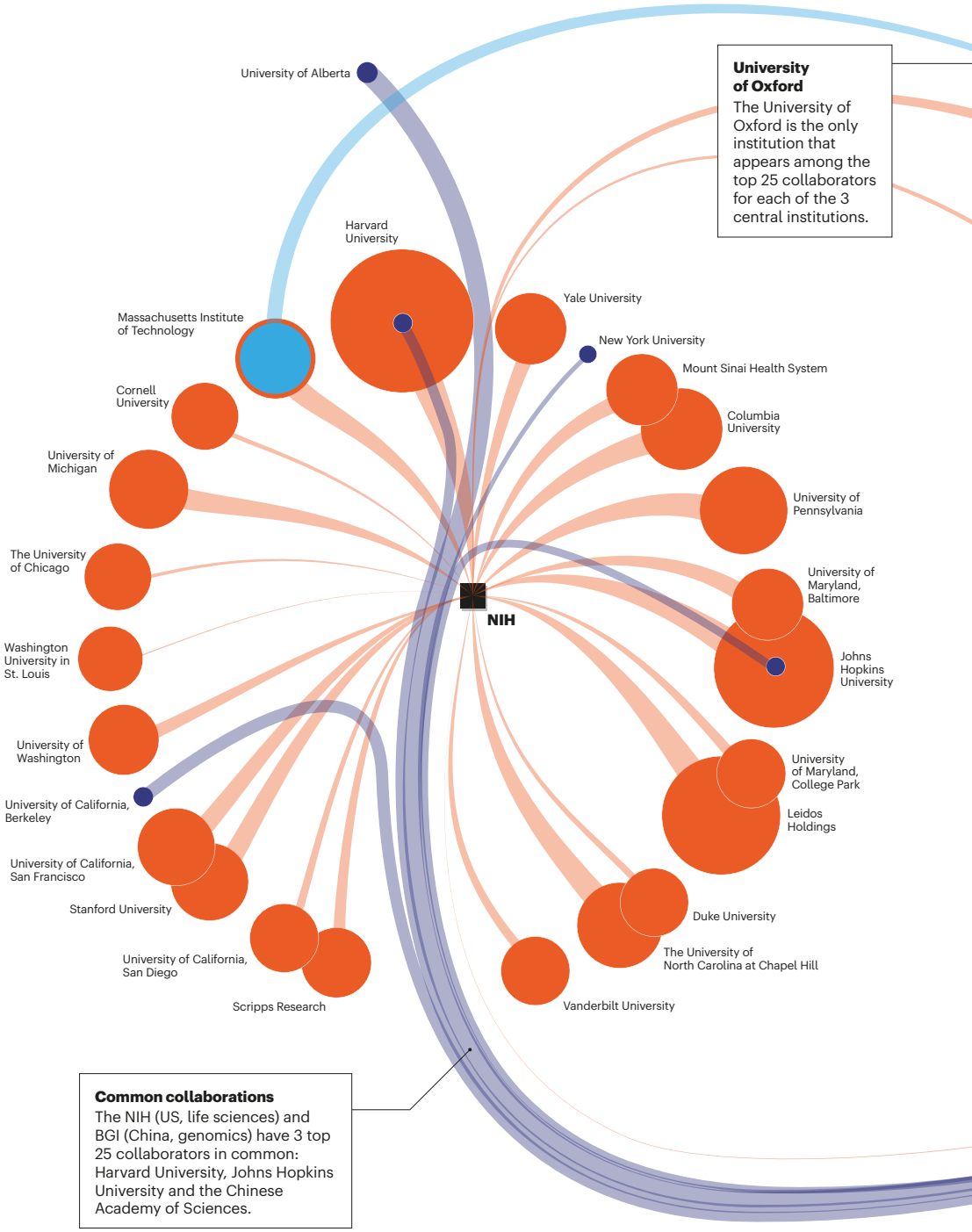
Data analysis by Bo Wu; infographic by Alisdair MacDonald

COLLABORATIVE CLUSTERS

The infographic shows the top 25 research partners of big science leading collaborators in 3 fields: high-energy physics, life sciences and genomics.

The Nature Index ranks institutions in the big science fields by their fractional counts (FC), referring to the share of their affiliated authors' contributions, and article counts (AC) in 82 high-quality journals. The table rankings (pages S39–S42) are for high affiliation articles only, meaning those with authors from 10 or more separate principal institutions.

The partner relationships shown are for the US National Institutes of Health (NIH), which ranks 2nd among the world's top institutions for producing big science research articles in the field of oncology and immunology (see page S42) and 3rd in the field of genetics (see page S39); the European Organization for Nuclear Research (CERN), in Switzerland, which is the 3rd biggest contributor to big science articles in physics and astronomy in the Nature Index (see page S40–S41); and BGI, a genome sequencing company that is China's biggest contributor to big science in genetics (see pS39). This infographic is based on all collaborative articles from the three institutions identified, regardless of the number of affiliations.

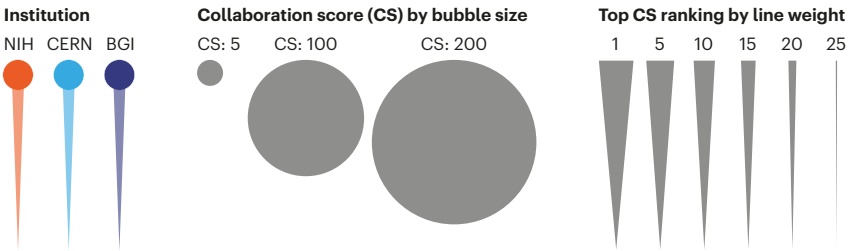


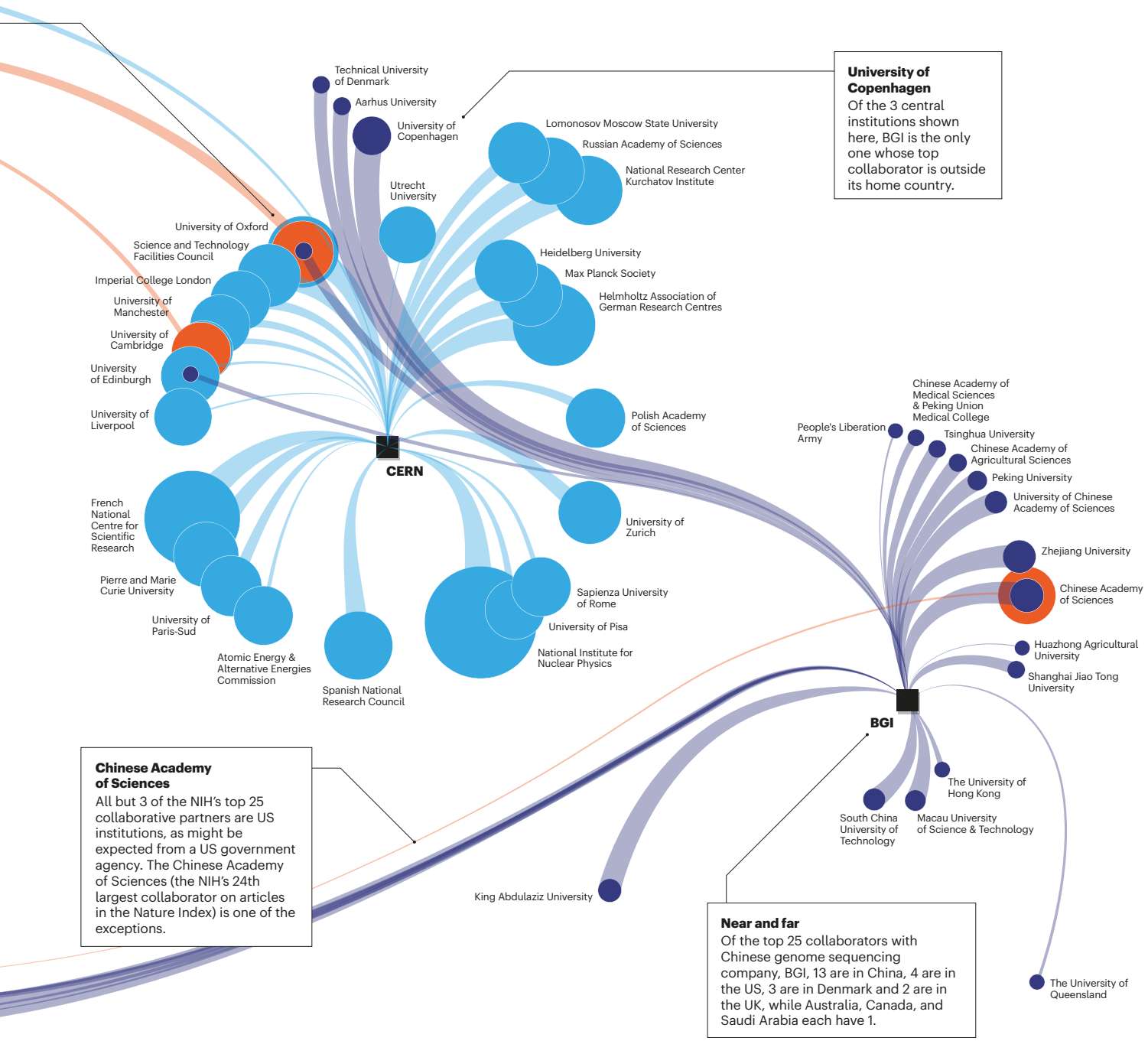
SOURCE: NATURE INDEX/DIMENSIONS FROM DIGITAL SCIENCE

LEGEND

The top 25 collaborators of the three central institutions are shown according to their joint collaboration score (CS) with the central institution, derived by summing the FCs* from articles with authors from both institutions. CS determines the size of the partner institutions' bubbles. The rank from 1 to 25 of their CS with the central institution is indicated by their line weight.

* For a definition of FC, see Collaborative clusters.





index

NIKLAS HALLE'N/AFP/GETTY IMAGES

Protesters at a central London rally highlight the perils for science if the UK leaves the European Union.

Brexit shadow hangs over EU partnerships

Uncertainty about the United Kingdom's role in EU science is damaging research networks. **By Mark Peplow**

More than three years after UK citizens narrowly voted for their country to leave the European Union, scientists still face great uncertainty about how Brexit will impact their research.

The United Kingdom has already missed three deadlines for leaving the EU. Although the terms of the country's withdrawal from the bloc have been agreed, political manoeuvring has stalled the ratification of this treaty. To break the deadlock, the United Kingdom is gearing up for a general election on 12 December, and faces a new Brexit deadline of 31 January 2020.

Even if the withdrawal treaty is ratified, it does little to resolve the country's long-term links with European science, which would be thrashed out in subsequent negotiations. If the United Kingdom leaves the EU next year without a deal in place to smooth the way, UK scientists could immediately lose access to the EU funding and collaborations that underpin their research.

This corrosive confusion is already reshaping collaborative networks. Some researchers are securing dual appointments that will enable them to straddle the United Kingdom and EU, and various UK universities have established partnership agreements with continental institutions so that their staff can continue to access EU funds.

"They want a Brexit-proof solution," says Kurt Deketelaere, secretary-general of the League of European Research Universities in Leuven, Belgium.

These changes could have far-reaching effects on existing partnerships. UK researchers collaborate more with researchers in other EU countries than any other region. According to the Nature Index, collaborative articles by UK–EU researchers in five leading journals (*Nature, Science, Proceedings of the National Academy of Sciences, Nature Communications* and *Science Advances*) grew by 36% between 2015 and 2018, although for reasons still not clear, the growth in collaborative articles

across all 82 journals tracked by the index has stalled since before 2016, when the vote to leave the EU was taken.

The United Kingdom has also benefited significantly from EU research programmes. Huge schemes such as Horizon 2020, the EU's €77 billion (US$85 billion) research and innovation programme, have provided a useful protocol for cross-border collaboration, facilitated by funding and helpful immigration arrangements. "That's a magic combination for doing research," says Graeme Reid, chair of science and research policy at University College London (UCL).

That success story is souring. In October, the Royal Society pointed out that the United Kingdom's annual share of EU research funding fell from €1.49 billion in 2015 to €1.06 billion in 2018, largely caused by a reduction in grant applications from UK researchers. It also found that the number of researchers coming to the United Kingdom through the EU's Marie Skłodowska Curie Fellowships has fallen by 35%, from 515 in 2015 to 336 in 2018.

Magic combination

"We have seen a dramatic drop in the number of leading researchers who want to come to the UK," says Venki Ramakrishnan, president of the Royal Society. "People do not want to gamble with their careers, when they have no sense of whether the UK will be willing and able to maintain its global scientific leadership."

Brexit comes at a crucial time for the future of EU science. The next research and innovation programme, Horizon Europe, will run from 2021 to 2027 and is expected to disburse about €100 billion. Almost half of this money is likely to go to large academia–industry collaborations in areas such as health, climate and food. The final budget and other details should be agreed by the end of 2020.

If Brexit goes ahead with a withdrawal agreement in place, most researchers hope it will pave the way for the United Kingdom to participate in Horizon Europe as an associate member, an option the UK government says it will consider. This might involve paying into the central fund, so that UK researchers can apply for grants in the same way as EU members, albeit with little influence over the programme's strategy.

Leaving without a deal would probably stymie the chances of associate membership altogether, and leave a huge question mark over the future status of the 17% of scientists working in the United Kingdom who are from other EU countries. It would also have a sudden impact on UK participation in ongoing EU collaborations, particularly affecting UK researchers responsible for the management and finance of EU projects.

Peter Coveney, a computational scientist at UCL, coordinates two major projects that apply advanced computing to biological modelling, backed by €12 million in Horizon 2020 funding. In September, he learnt that the European Commission would ask UK project coordinators to step down in the event of a no-deal Brexit. To counter that, Coveney is ready to move the management of his projects to the continent. In March, he accepted a professorship in applied high-performance computing at the University of Amsterdam in the Netherlands, which he holds in addition to his existing UCL roles. "It's a Brexit mitigation strategy," he says.

Leading UK universities are adopting a similar approach, signing cooperation agreements with partners in continental Europe to intensify collaboration, establish joint research programmes and exchange staff and students. The University of Cambridge has partnered with the Ludwig Maximilians University of Munich in Germany, for example, and the University of Oxford has established an office in Berlin to facilitate partnerships with institutions there.

Some regions of the United Kingdom are particularly vulnerable to Brexit, and may struggle to adapt. Northern Ireland, for example, depends heavily on collaborations with partners from the Republic of Ireland, a separate country and EU member state. "Sixty-three per cent of Horizon 2020 applications from researchers in Northern Ireland involved a partner from the republic," says Gerry McKenna, who chairs the Royal Irish Academy's North–South Committee, which is concerned with collaboration across the Irish border. "There's no way that Brexit won't have some negative impact on research collaboration."

"There's no way that Brexit won't have some negative impact on research collaboration."

Amid these damage limitation efforts, Reid has been looking farther afield for ways to diversify the United Kingdom's research collaborations. He and statistician, Adrian Smith, who leads the London-based Alan Turing Institute, have written a report for the UK government that outlines opportunities for international collaboration if the country decides not to associate with Horizon Europe. Published on 5 November, the report suggests measures such as dedicated funding streams to enhance global collaboration, and new fellowships to attract talent to the United Kingdom. Reid hopes that the report will demonstrate that there are alternatives to traditional EU-focused collaborations.

The UK government has also tried to reassure the domestic research community with a string of policy announcements, such as reaffirming a commitment to increase national spending on research and development to at least 2.4% of GDP by 2027.

It also said that international students will be allowed to remain in the country for up to two years after graduation, so they can seek employment in the United Kingdom, and promised fast-track visa routes for foreign researchers. "I think that's an enormous step forward," says Martin Smith, a policy manager at Wellcome.

Fast track

However, the UK government's rhetoric has focused on attracting 'the brightest and best' from abroad. Smith says the new visa systems must cover all levels of the scientific workforce, including lab technicians and postdoctoral researchers, and adds that the costs of the system will be a key issue for early-career researchers.

If there is a no-deal Brexit, the UK government has promised that it will honour funding for all successful competitive UK bids to Horizon 2020 up to the end of next year, potentially costing hundreds of millions of pounds. But this overlooks the contribution of other EU funding streams.

EU structural funds, for example, are used to boost the economic development of EU regions that may be lagging behind, by investing in projects that can enhance innovation and create jobs.

"Structural funds have been crucial in building up the research base in Northern Ireland to an internationally competitive level," says McKenna. Several research facilities in Northern Ireland depended on such structural funding, he says, including the Northern Ireland Science Park.

Funding aside, the United Kingdom may simply become less attractive to international researchers if it adopts a hostile stance towards the EU during future negotiations. "This is not just a financial equation," says Reid. "An environment that is welcoming and nurturing is critical."

Deketelaere says that top researchers are already choosing to take their research out of the country. "In continental Europe, we're seeing an enormous influx of unsolicited applications for jobs in our universities, from excellent people," he says. This reshaping of the research landscape could become Brexit's lasting legacy for science.

Mark Peplow is a science writer based in Cambridge, UK.

index

Conducting the multi-author choir

Large research teams can produce higher impact work than scientists who go it alone, but organizing a paper produced by multitudes can be a major challenge. **By Jack Leeming**

IAN POTTEN/BRITISH ANTARCTIC SURVEY

Andrew Shepherd and his team collect cores from the George VI Ice Shelf, Antarctica, in their work estimating ice loss from polar caps.

For a frog, exposure to the amphibian chytrid fungus (*Batrachochytrium dendrobatidis*) is very bad news indeed. The fungus thrives in the same wet, hot conditions that frogs favour and it grows on amphibian skin. Frogs breathe through their skin, which is used by almost all species for electrolyte exchange. Chytrid prevents electrolytes from entering the animal's body, which eventually causes a heart attack.

Chytrid fungus species are responsible for significant amphibian population reductions in Central and North America, Europe and Australia. Although declines were at their worst in the 1980s, one 2004 study suggested that at least 43% of amphibian species are dwindling worldwide. New Guinea, home to 6% of the world's frog species, is one place chytrid is yet to invade.

Deborah Bower, an ecologist at the University of New England in Armidale, Australia, is investigating proactive protection strategies for New Guinea, including increased quarantine measures and an island-wide surveillance programme. In 2015, she collaborated with 29 other scientists on these and other recommendations. The results were published in June 2019 in *Frontiers in Ecology and the Environment*.

Such collaboration is unusual in Bower's field, where single-author papers are common. "When the fungus gets to New Guinea, more than 100 frog species could go extinct," she says. "The island has a complex political system; it's half Papua New Guinea and half Indonesia. There's not much local experience in dealing with the disease. We brought in scientists from the US and Australia who had experience with chytrid, plus experts from a policy background who have worked with governments on large-scale changes."

Bower hopes the paper will have more impact than something she could have produced alone. "We had authors from five different countries, which gave us access to more skills. One of the co-authors did climatic modelling on the fungus, for example."

Another bonus, adds Bower, is a more refined document. "There are more eyes going over it," she says.

At the same time, this presented a challenge. Organizing schedules was difficult, almost as much as fielding multiple pages of feedback. "Getting comments from 29 people is overwhelming. I had maybe 20 different documents to go through. A colleague travelled 460 kilometres from Sydney and we sat down with two computers, one with tracked changes and one with comments, to manage it."

Another struggle was editorial policies and style guides. In 2017, *Science* accepted a 'perspective' article from Bower and four co-authors to outline work in progress. "I had to ditch 25 authors," she says, noting that the journal sets a limit of five authors for its perspective articles.

Greater impact

Scientists are co-authoring more than ever before. In 2016, *The Economist* reviewed more than 34 million research papers published between 1996 and 2015, and found that the average author numbers grew from 3.2 to 4.4 per paper. A 2018 Nature Index analysis found that the field of high-energy physics is largely responsible for the rise of papers authored by more than 1,000 individual scientists in recent years. A 2014 collaboration estimating the size of the Higgs boson, for example, listed a record-breaking 5,154 authors.

Papers in physics and astronomy, genetics, oncology and immunology, were also identified by Nature Index to be the most likely to have long author lists.

CERN's particle-collision experiment, ATLAS, which is designed to test the standard model of physics, makes use of one of two general-purpose detectors at the Large Hadron Collider in Switzerland, and regularly produces physics megapapers. Karl Jakobs, a physicist and ATLAS collaboration spokesperson, is part of a publishing group that makes up the 3,000-strong team. "In terms of who actually writes the paper, we assign two or three editors, called an editorial team," says Jakobs. "They discuss, present an outline and write the paper, with input from the scientists who provided the data."

> **"The main challenge is some very large characters in the project, who have very different opinions."**

Papers are then subject to a series of internal peer reviews and the editorial team takes in contributor feedback, followed by an external institutional review from a collaborating physics department. The extent of this back-and-forth editing before a paper is even submitted to a journal might seem intimidating for some, but "it's not a nightmare", says Jakobs, because the collaboration and review process is well defined and organized in a way that makes sense to the participants.

Big personalities

The ATLAS collaboration published more than 100 papers last year, but similar efforts can be a struggle in other fields. Andrew Shepherd, an Earth observation scientist at the University of Leeds, UK, leads a publishing consortium of 96 researchers who estimate sea-ice loss from polar ice caps.

"In 2010, when we started working on this, there were probably around 50 individual estimates in the literature for how much ice was being lost," Shepherd says. "The project was established to shed light on why there were such large differences between individual estimates, and then to produce a single estimate for the community.

"The main challenge is some very large characters in the project who have very different opinions as to whether ice is being lost or gained, for instance," he says.

Shepherd explains that organizing the input of almost 100 experts on a topic of great interest to the press and public can be difficult. "It's very intensive on my part. I calculated that, in the first assessment, I sent 5,000 e-mails on the project. We've done three assessments now. I haven't counted those e-mails because it's quite depressing, but it's probably about the same," he says.

The ice sheet mass balance inter-comparison exercise (IMBIE), which has produced high-profile papers in *Science* (2012) and *Nature* (2018), "ate into my summer holidays", Shepherd admits. "I've spent probably about two hours every morning of my vacation this year dealing with the minutiae of publishing this most recent paper. It demands a lot of time."

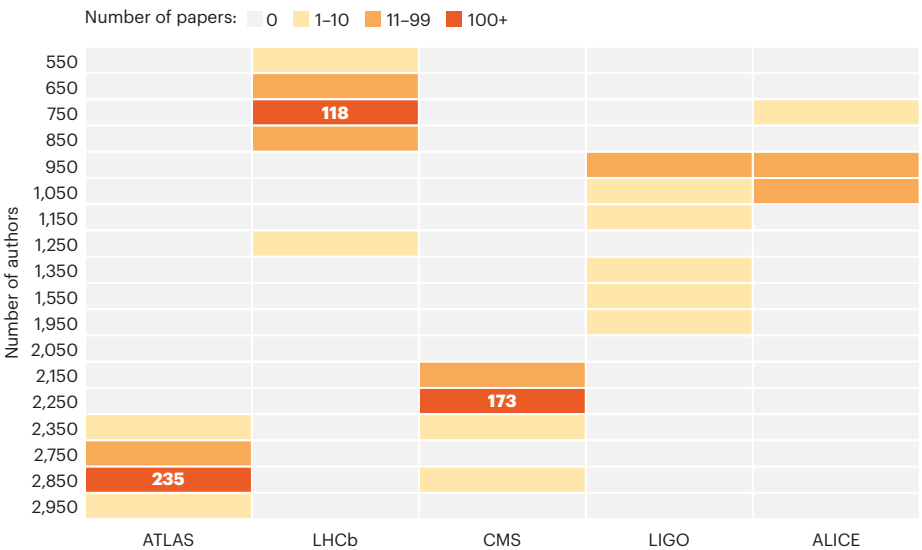
Shepherd says this time investment has its rewards. The group's first publication estimated sea-ice loss of around 4,000 gigatonnes from the Antarctic and Greenland ice sheets since 1992. The paper has clocked more than 800 citations since 2012, which is "very, very high for the climate sciences", Shepherd says. It also has a high Altmetrics score, which measures interest from social media and online news audiences.

"We have a broader impact," says Shepherd. "The US Environmental Protection Agency has been using our data for the past four or five years as a climate indicator, for example. It's a very rewarding project."

Jack Leeming is an editor for *Nature* based in London.

BIG PHYSICS PUBLICATIONS PATTERNS

The heat map shows the quantity of research papers according to the number of authors published by 4 large collaborative physics experiments at CERN in Switzerland (ATLAS, LHCb, CMS and ALICE) as well as the Laser Interferometer Gravitational-Wave Observatory (LIGO) in the US.



SOURCE: NATURE INDEX / DIMENSIONS FROM DIGITAL SCIENCE